DOCUMENT RESUME

ED 470 320 PS 030 834

AUTHOR	Rock, Donald A.; Pollack, Judith M.
TITLE	Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K): Psychometric Report for Kindergarten through First Grade. Working Paper Series.
INSTITUTION	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-WP-2002-05
PUB DATE	2002-08-00
NOTE	199p.; Prepared with contributions by Sally Atkins-Burnett, Tom Hoffer, Samuel J. Meisels, Karen Tourangeau, Jerry West, and Elvira Germino Hausken. For the ECLS-K CD-ROM, see ED 463 872.
AVAILABLE FROM	ED Pubs, P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827 (Toll Free); Fax: 301-470-1244; e-mail: edpubs@inet.ed.gov. For full text: http://nces.ed.gov/pubsearch.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC08 Plus Postage.
DESCRIPTORS	Academic Achievement; Achievement Tests; Catholic Schools; *Cognitive Ability; *Cognitive Measurement; Comparative Analysis; *Elementary School Students; Interpersonal Competence; Item Response Theory; *Kindergarten Children; Longitudinal Studies; Mathematics Achievement; Measurement Techniques; Measures (Individuals); Primary Education; Private Schools; *Psychometrics; Psychomotor Skills; Public Schools; Reading Achievement; Research Problems; Sex Differences; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS	*Early Childhood Longitudinal Survey

ABSTRACT

The Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), selected a nationally representative sample of approximately 22,000 kindergartners in the fall of 1998 and is following these children through the end of the fifth grade. Baseline data about these children, their families, and their kindergarten programs were collected by means of telephone interviews with the children's parents or guardians and from self-administered questionnaires completed by the kindergarten teachers. Data were also gathered during an individual assessment with each child. This report documents the design, development, and psychometric characteristics of the assessment instruments used in the ECLS-K. The focus is on the psychometric results of the assessment instruments for four time points: Fall- and Spring-kindergarten and Fall- and Spring-first grade. The assessment instrument examined three domains: the cognitive (direct and indirect), socioemotional, and psychomotor. In addition, the report discusses issues involved in analyzing longitudinal measures of cognitive skills, including the use of total scores and of proficiency probabilities to measure longitudinal change. Initial results revealed sex differences in prereading skills at kindergarten entry and the areas of gain. Public school children had the lowest reading skills at kindergarten entry, followed by Catholic school children, with private non-Catholic school children having the highest

Reproductions supplied by EDRS are the best that can be made
from the original document.

reading skills. There were differences in the areas of gain in children attending different types of schools. The report's five appendices include a summary of national mathematics and science curriculum standards, reading assessment content classifications used for test item development, ECLS item parameters and item fit by rounds, and score statistics for indirect and psychomotor measures for selected subgroups. (Contains 60 references.) (KB)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PS

ED 470 320

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade

Working Paper No. 2002–05 August 2002

Contact: Elvira Germino Hausken
Early Childhood, International and Crosscutting Studies Division
elvira.hausken@ed.gov

PS 030834

U. S. Department of Education
Office of Educational Research and Improvement

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

The Working Paper Series was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

U. S. Department of Education
Office of Educational Research and Improvement

Foreword

In addition to official NCES publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The ***Working Paper Series*** was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/), or contact Sheilah Jupiter at (202) 502–7444,
e-mail: sheilah.jupiter@ed.gov, or mail: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 1990 K Street NW, Room 9048, Washington, DC 20006.

Marilyn M. Seastrom
Chief Mathematical Statistician
Statistical Standards Program

Ralph Lee
Mathematical Statistician
Statistical Standards Program

Preface

The Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K) is an ongoing study by the U.S. Department of Education, National Center for Education Statistics (NCES) that focuses on children's early school experiences beginning with kindergarten. The study follows a nationally representative sample of approximately 22,000 children from kindergarten through fifth grade. Four rounds of data have been collected; fall of 1998, spring of 1999, fall of 1999, and spring of 2000. Additional spring follow-up data collections are scheduled for 2002 and 2004. The ECLS-K is conducted under the sponsorship of the National Center for Education Statistics, (NCES), with additional funding and technical support from the Office of Special Education Programs, Office of the Under Secretary Planning and Evaluation Service, and the Office of Bilingual Education and Minority Languages Affairs of the U.S. Department of Education, the Economic Research Service of the U.S. Department of Agriculture, the National Institute of Child Health and Human Development of the National Institutes of Health, and the Administration on Children, Youth and Families, Head Start Bureau of the U.S. Department of Health and Human Services.

Because of the magnitude and complexity of the ECLS-K, NCES and its contractor, Westat, continue to design and test the instruments that will be used in the 2002 and 2004 spring follow-up data collections. This paper is one of several that have been prepared in support of the ECLS-K design activities. While the information and recommendations found in this paper have contributed to the design of the ECLS-K, specific methods and procedures may or may not actually be incorporated into the final ECLS-K design. It is our hope that the information found in this paper will not only provide background for the development of the ECLS-K, but that it will be useful to researchers developing their own studies of young children, their families, and their educational experiences.

Jerry West
Program Director
Early Childhood Longitudinal Studies

Val Plisko
Associate Commissioner
Early Childhood, International and Crosscutting Studies Division

Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade

Prepared by:

Donald A. Rock and Judith M. Pollack
Educational Testing Service

With Contributions By:

Sally Atkins-Burnett, University of Michigan
Tom Hoffer, NORC
Samuel J. Meisels, University of Michigan
Karen Tourangeau, Westat
Jerry West, NCES
Elvira Germino Hausken, NCES

Prepared for:

U.S. Department of Education
Office of Educational Research and Improvement
National Center for Education Statistics

August 2002

TABLE OF CONTENTS

Chapter Page

1 INTRODUCTION 1-1

2 DESIGN OF THE ASSESSMENT INSTRUMENTS 2-1

2.1 Development of Cognitive Test Specifications: Domains 2-1
2.2 Direct Cognitive Test 2-2

2.2.1 Individually Administered Adaptive Tests 2-2
2.2.2 Sources of the ECLS–K Frameworks 2-3
2.2.3 Item and Time Allocations 2-6
2.2.4 Mathematics Test Specifications 2-7
2.2.5 Reading Test Specifications 2-10
2.2.6 General Knowledge: Science and Social Studies Test Specifications 2-13

2.3 Indirect Cognitive Assessment: Academic Rating Scale 2-17
2.4 Social Rating Scales: Teacher and Parent 2-20
2.5 Psychomotor Assessment 2-21
2.6 Oral Language Development Scale 2-21

3 DEVELOPMENT OF THE TWO-STAGE DIRECT TEST FORMS 3-1

3.1 Development of the Item Pool 3-1
3.2 Field Testing and Item Analysis 3-1

3.2.1 Differential Item Functioning Analysis 3-3
3.2.2 Field Test Conclusions 3-4

3.3 Assembly of the Final Adaptive Forms 3-4

3.3.1 Two-Stage Testing Procedure 3-4
3.3.2 Criterion-Referenced Item Clusters 3-9
3.3.3 English Fluency and Spanish Mathematics Test 3-11

4 ITEM RESPONSE THEORY SCALING FOR LONGITUDINAL MEASUREMENT AND EQUATING TO EARLIER ROUNDS 4-1

4.1 Overview of Item Response Theory 4-1
4.2 Item Response Theory Estimation Using PARSCALE 4-5

TABLE OF CONTENTS (continued)

Chapter Page

5 PSYCHOMETRIC CHARACTERISTICS OF THE ECLS–K DIRECT COGNITIVE BATTERY 5-1

5.1 Motivation and Timing 5-1
5.2 Differential Item Functioning 5-4
5.3 Reading Test 5-5

5.3.1 Samples and Operating Characteristics 5-5
5.3.2 Reliabilities 5-7
5.3.3 Score Gains 5-9
5.3.4 Differential Item Functioning 5-9

5.4 Mathematics Test 5-11

5.4.1 Samples and Operating Characteristics 5-11
5.4.2 Reliabilities 5-12
5.4.3 Score Gains 5-14
5.4.4 Differential Item Functioning 5-14
5.4.5 Comparability of Spanish Mathematics Test 5-15

5.5 General Knowledge Test 5-17

5.5.1 Samples and Operating Characteristics 5-17
5.5.2 Reliabilities 5-18
5.5.3 Score Gains 5-19
5.5.4 Differential Item Functioning 5-19

5.6 Intercorrelations of the Direct Cognitive Measures within Rounds 1 to 4 5-20
5.7 Test Results by Round and Selected Demographics 5-21
5.8 Test Item Usage and Item Performance 5-21
5.9 Interviewer Variance as a Threat to Validity 5-38

6 PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT AND PSYCHOMOTOR MEASURES 6-1

6.1 Indirect Cognitive Assessment Using the Academic Rating Scale 6-2
6.2 Item Response Theory 6-3

6.2.1 One Parameter Item Response Theory 6-3
6.2.2 Item Response Theory Estimation Using Winsteps 6-4
6.2.3 Floor and Ceiling 6-8

TABLE OF CONTENTS (continued)

Chapter | Page

6.3 Social Rating Scale 6-8
6.4 Psychomotor Assessment 6-10
6.5 Discriminant and Convergent Validity of the Direct and Indirect Measures 6-11

7 PSYCHOMETRIC CHARACTERISTICS OF THE ENGLISH AND SPANISH ORAL LANGUAGE DEVELOPMENT SCALE 7-1

8 APPROACHES TO MEASURING CHANGE USING ECLS–K LONGITUDINAL SCORES 8-1

8.1 Total Scores to Measure Longitudinal Change 8-2
8.2 Proficiency Probabilities to Measure Longitudinal Change 8-3
8.3 Method 8-4
8.4 Results 8-5

8.4.1 Gender and Location of Maximum Gain 8-5
8.4.2 School Sector and Location of Maximum Gain 8-9

8.5 Alternative Measure of Overall Gain 8-15
8.6 Conclusions 8-16

REFERENCES R-1

List of Appendixes

Appendix

A Summary of 1989 NCTM Mathematics Curriculum Standards for Grades Kindergarten to Fourth A-1

B Reading Assessment Detailed Content Classifications Used for Item Development B-1

C Summary of the 1996 National Research Council Grades Kindergarten to Fourth Content Standards in Science C-1

D ECLS Item Parameters and Item Fit by Rounds D-1

E Score Statistics for Indirect and Psychomotor Measures for Selected Subgroups E-1

TABLE OF CONTENTS (continued)

List of Tables

Table		Page
2-1	Proposed mathematics longitudinal test specifications, in percentages of testing time, for kindergarten through fifth grade	2-9
2-2	Proposed reading longitudinal test specifications, in percentage of testing time, for kindergarten through fifth grade	2-12
2-3	ECLS-K science longitudinal test specifications, in percentages of testing time, for kindergarten through fifth grade	2-14
2-4	Proposed social studies longitudinal test specifications, in percentages of testing time, for kindergarten through first grade	2-16
2-5	Academic rating scale response scale	2-19
2-6	Social rating scale response scale	2-20
3-1	Routing cutting score and anticipated second-stage percentages	3-8
5-1	Child's overall motivation level during the assessment	5-2
5-2	Child's overall cooperation during the assessment	5-3
5-3	Child's overall attention level during the assessment	5-3
5-4A	Reading test: samples and operating characteristics	5-6
5-4B	Reading test: reliabilities and mean score gains	5-8
5-4C	Reading test: differential item functioning	5-10
5-5A	Mathematics test: samples and operating characteristics	5-12
5-5B	Mathematics test: reliabilities and mean score gains	5-13
5-5C	Mathematics test: differential item functioning	5-14
5-5D	Performance of children who took Spanish mathematics test in round 2 and English mathematics test in round 4	5-17
5-6A	General knowledge test: samples and operating characteristics	5-18

5-6B General knowledge test: reliabilities and mean score gains 5-19

5-6C General knowledge test: differential item functioning 5-20

5-7 Intercorrelations of the direct cognitive measures within rounds 1 to 4 5-18

5-8 Reading Item Response Theory theta score (range of possible values: -5 to 5) 5-20

5-9 Mathematics Item Response Theory theta score (range of possible values: -5 to 5) 5-23

5-10 General knowledge Item Response Theory theta score (range of possible values: -5 to 5) 5-24

5-11 Reading Item Response Theory scale score (range of possible values: 0 to 92) 5-25

5-12 Mathematics Item Response Theory scale score (range of possible values: 0 to 64) 5-26

5-13 General knowledge Item Response Theory scale score (range of possible values: 0 to 51) 5-27

5-14 Probability of proficiency, reading level 1: letter recognition (range of possible values: 0.0 to 1.0) 5-28

5-15 Probability of proficiency, reading level 2: beginning sounds (range of possible values: 0.0 to 1.0) 5-29

5-16 Probability of proficiency, reading level 3: ending sounds (range of possible values: 0.0 to 1.0) 5-30

5-17 Probability of proficiency, reading level 4: sight words (range of possible values: 0.0 to 1.0) 5-31

5-18 Probability of proficiency, reading level 5: beginning comprehension (range of possible values: 0.0 to 1.0) 5-32

5-19 Probability of proficiency, mathematics level 1: number and shape (range of possible values: 0.0 to 1.0) 5-33

5-20 Probability of proficiency, mathematics level 2: relative size, etc. (range of possible values: 0.0 to 1.0) 5-34

TABLE OF CONTENTS (continued)

List of Tables (continued)

Table		Page
5-21	Probability of proficiency, mathematics level 3: number sequence, etc. (range of possible values: 0.0 to 1.0)	5-35
5-22	Probability of proficiency, mathematics level 4: addition/subtraction (range of possible values: 0.0 to 1.0)	5-36
5-23	Probability of proficiency, mathematics level 5: multiplication/division (range of possible values: 0.0 to 1.0)	5-37
5-24	Components of variance associated with the child, interviewer, and team leader for fall-kindergarten	5-38
5-25	Components of variance associated with the child, interviewer, and team leader for spring-first grade	5-38
6-1	Person reliability for the Rasch-based score	6-5
6-2	Fit statistics for Persons and Items	6-6
6-3	Academic rating scale means and standard deviations (range of possible values: 1 to 5)	6-7
6-4	Percent of sample with perfect and minimum academic rating scale scores in kindergarten and first grade	6-8
6-5	Teacher social rating scales: split half reliability	6-9
6-6	Parent social rating scales: split half reliability	6-9
6-7	Teacher social rating scales: means and standard deviations (range of possible values: 1 to 4)	6-10
6-8	Parent social rating scales: means and standard deviations (range of possible values: 1 to 4)	6-10
6-9	Psychomotor scales: means and standard deviations (round 1 only)	6-10
6-10	Intercorrelations among the indirect cognitive teacher ratings, selected teacher and parent socio-behavioral measures, and direct cognitive test scores	6-12

TABLE OF CONTENTS (continued)

List of Tables (continued)

Table Page

7-1 English Oral Language Development Scale test results 7-2

7-2 Spanish Oral Language Development Scale test results 7-3

8-1 Means, standard deviations, and correlations of reading scale scores fall- and spring-kindergarten 8-2

8-2 Gender multilevel binomial analysis of gains at levels 4 to 5 versus gains at levels 1 to 3 8-8

8-3 School sector binomial multilevel analysis of gains at levels 4 to 5 versus gains at levels 1 to 3 8-12

8-4 Multilevel analysis of raw gains by school sector and location of maximum gain (locmg) 8-14

List of Figures

Figure

3-1 Two-stage adaptive design 3-5

4-1 Probability of correct answer 4-2

4-2 Items with different difficulty (b) 4-4

4-3 Items with different discrimination (a) 4-4

8-1 Proficiency levels theta scale 8-4

8-2 Fall to spring reading gains by gender, adjusted for age and parent education 8-6

8-3 Location of maximum gain by gender, fall- to spring-kindergarten 8-7

8-4 Adjusted mean changes on the total score scale by gender and proficiency level 8-9

8-5 Fall to spring reading gains by school sector, adjusted for age and parent education 8-10

TABLE OF CONTENTS (continued)

List of Figures (continued)

Figure Page

8-6 Location of maximum reading gain by school type, fall- to spring-kindergarten 8-11

8-7 Adjusted mean gains on the total score scale by school sector and location of maximum gain 8-13

1. INTRODUCTION

This report documents the design, development, and psychometric characteristics of the assessment instruments used in the Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K). The ECLS–K is sponsored by the U.S. Department of Education, National Center for Education Statistics. The ECLS–K was designed to assess the relationship between a child's academic and social development and a wide range of family, school, and community variables. While the ECLS–K will ultimately span kindergarten through fifth grade, this report documents the psychometric results for four time points—fall- and spring-kindergarten and fall- and spring-first grade.

Three domains are represented by the ECLS–K kindergarten and first grade (K–1) assessment instruments: cognitive (direct and indirect), socioemotional, and psychomotor. Direct cognitive measures refer to scores based on children's "direct" responses to cognitive test items. In kindergarten and first grade, direct cognitive tests were administered in reading, mathematics, and general knowledge. Indirect cognitive measures were ratings by teachers of the children's cognitive performance in closely related areas: language and literacy, mathematical thinking, and general knowledge. The socioemotional measures are ratings by parents and teachers of children's social skills and approaches to learning. The psychomotor assessment tested children's fine and gross motor skills.

The direct cognitive assessments for kindergarten and first grade were designed to measure an individual child's knowledge at a given point in time, as well as to measure that same child's academic growth on a vertical score scale based on successive assessments. In addition to designing the cognitive tests to make reliable normative comparisons with respect to status and growth, the tests were also designed to provide criterion-referenced interpretations. That is, in the reading and mathematics content domains, criterion-referenced proficiency scores can be used to describe a given child's mastery of specific knowledges that mark ascending critical points on the developmental growth curve. These multiple criterion-referenced levels serve two functions. First, they help with respect to the interpretation of what a particular attained score level means in terms of what a child can or cannot do. Second, they are useful in measuring change at particular score points along the score scale. Thus they provide a means of evaluating the influence of children's experiences on changes in mastery of specific skills.

The development of the direct cognitive battery was carried out in five steps:

1. A background review was carried out using all the currently available psychometric instruments and the constructs that they purported to measure.
2. Tests specifications were developed that were appropriate for the domains and constructs considered relevant for kindergarten through first grade.
3. An item pool was developed that reflected the test specifications in step 2.
4. The item pool was field tested in order to gather statistical and psychometric evidence as to the appropriateness of the items for carrying out the overall assessment goals.
5. The final test forms were assembled consistent with field test item statistics and the test specifications.

Chapter 2 of this report describes the objectives and design of the assessment instruments. For the cognitive tests, this includes selection of content domains, and for the direct cognitive tests, the rationale for individually administered adaptive tests, the source, and development of frameworks. Chapter 3 describes the development and field testing of the item pools for the direct cognitive measures, the selection of test items for the final forms, and the creation of a Spanish-language version of the mathematics assessment. It also introduces the criterion-referenced subsets of items selected for the reading and mathematics tests. Chapter 4 contains an overview of the Item Response Theory (IRT) procedures used to scale the test scores. Chapter 5 presents the psychometric characteristics of the direct cognitive tests given in kindergarten and first grade. Chapter 6 describes the development and psychometric characteristics of the indirect and psychomotor measures and discusses the relationship between the direct and indirect measures of the cognitive domains. Chapter 7 describes performance on the Oral Language Development Scale (OLDS), the instrument used to evaluate children's fluency in English and Spanish. Chapter 8 presents issues in analyzing longitudinal measures of cognitive skills.

Some children's English language fluency was not sufficient for them to participate in the direct cognitive assessment. Those who spoke Spanish received a Spanish translation of the mathematics test and psychomotor assessment. By spring of first grade, more than two-thirds of these Spanish-speaking children had developed enough fluency in English to receive the English version of the test. Chapter 2 describes the selection of the language assessment measures, chapter 3 provides background on the development of the Spanish mathematics test form, and chapter 5 presents evidence supporting the comparability of the scores derived from the English and Spanish versions. Speakers of other languages did not participate in the direct cognitive assessment (and Spanish-speakers in the reading and general knowledge sections) until their English language skills were sufficiently developed to take the ECLS–K K–1 tests in English.

A national probability sample of about 22,000 children in about 800 public and 200 private schools were assessed at entry to kindergarten in fall 1998 (round 1). They were followed up in spring-kindergarten (round 2) and fall- and spring-first grade (rounds 3 and 4, respectively). The third round (fall-first grade) was a subsample of about 30 percent of the longitudinal cohort. The direct cognitive assessments were conducted in all four rounds of data collection, while the indirect cognitive and socioemotional measures were collected in rounds 1, 2, and 4 (fall- and spring-kindergarten, and spring-first grade). The psychomotor assessment was administered only in round 1 (fall-kindergarten).

Sample counts, completion rates, and breakdowns by gender, race/ethnicity, socioeconomic scale, and school type are presented in the psychometric analyses in Chapter 5 and Appendix E. Additional information about the sample design, the assessment instruments, and the collection of assessment data can be found in the ECLS–K Electronic Code Books and user manuals.[1]

---

[1] *ECLS–K Restricted-Use Base Year User's Manual* (NCES 2000–097), August 2000; *ECLS–K Base Year Public-Use User's Manual* (NCES 2001–029), October 2000; *ECLS–K First Grade Restricted-Use Electronic Code Book* (NCES 2001–128), November 2001; *ECLS–K First Grade Restricted-Use User's Manual* (NCES 2002–189), November 2001.

2. DESIGN OF THE ASSESSMENT INSTRUMENTS

The ECLS–K was designed to assess children's academic and social development during the kindergarten through fifth grade years. Direct and indirect cognitive measures describe children's academic performance at each time point, as well as measuring growth over time. Measures of children's social behaviors and approaches to learning are reported in the social rating scales derived from teachers' and parents' observations in the school and home settings. The psychomotor assessment scales measuring fine and gross motor skills at kindergarten add contextual information specific to each child. The cognitive and social skills measures, along with contextual variables in the ECLS–K database collected from schools, parents, teachers, and children provide a basis for studying the relationships between a child's academic and social development and a wide range of family, school, and community variables. Analysis of these assessment scores can provide the basis for policy-relevant analysis of growth rates, school influences, and subgroup differences in achievement and growth.

The National Center for Education Statistics (NCES) and contractor staff assembled school curriculum specialists, teachers, and academicians to consult on the design and development of the assessment instruments. Topics addressed included domains to be covered, test specifications, mode of administration, and time allocations. The advice of these experts guided the decisions necessary to make efficient use of resources while minimizing burden on teachers and students.

2.1 Development of Cognitive Test Specifications: Domains

The panel of experts recommended that the knowledge and skills assessed by the ECLS–K tests should represent the typical and important cognitive goals of elementary schools' curricula. The subject-matter domains of language use and literacy skills (referred to hereafter simply as "reading" for the direct cognitive assessment), mathematics, and general knowledge (science and social studies) were selected. This focus on the main academic subjects of the elementary grades came about because of the central nature of these skills as being the antecedents of individuals' later educational outcomes. The practical difficulties of adequately assessing children's proficiencies in writing, art, and music within the resource constraints of the study precluded assessment in these domains.

2.2 Direct Cognitive Test

The nature of the ECLS–K cognitive assessment battery was shaped by its basic objectives and constraints. Foremost among these was the requirement that the test battery accurately measure children's cognitive development across the whole span of the study. The longitudinal design of the study required that a vertical scale (one on which the scores of kindergartners to fifth graders can be placed) in each subject area be developed that can support the measure of valid change scores. Such a scale would allow one to compare achievement levels across grades and to quantify the gains children make from year to year. The goal of minimizing time, cost, and burden on students and teachers shaped the kinds of test items that could be used, as well as the structure of the tests. On average, the amount of time to test each child in all three domains was about 50 minutes in each assessment cycle. This limitation precluded the use of assessment tasks such as extended reading passages or hands-on science experiments.

2.2.1 Individually Administered Adaptive Tests

During the background review, the contractor staff, which included experts in child development, primary education, and testing methodology, in collaboration with their counterparts at NCES, made the recommendation that the direct cognitive measures be administered individually to each sampled child. Since young children are not experienced test-takers, individual administration could provide more sensitivity to each child's needs than a group-administered test. In addition to being individually administered, it was also recommended that the tests be adaptive in nature; that is, each child should be tested with a set of items that is most appropriate for his or her level of achievement.[2]

The development of a vertical scale that must span kindergarten to fifth grade and have optimal measurement properties throughout the achievement range calls for multiple test forms that vary in their level of difficulty. Although the forms are tailored for individuals within a grade, the overall grade-level forms should reflect core curriculum elements for that particular grade. At the same time there must be overlapping items in forms within a grade, as well as across grades. These linking items tie the vertical scale together both across forms within a grade and across grades. About 20 to 30 percent of the items should overlap between adjacent grades.

---

[2] The ECLS-K assessments are not timed tests, so students can take as much time as necessary to complete them.

A child who is essentially performing on grade level should receive items that span the curriculum for his or her grade. Children whose achievement is above or below grade level should be given tasks with difficulty levels that match their individual level of development at the time of testing, rather than a grade-level standard. A child who is performing much better in relation to his or her cohorts, as measured by a brief routing test, would subsequently be given test items that are proportionately more difficult, while a child performing below grade level would receive a form with proportionately more easy items. The matching of the difficulties of the item tasks to each child's level of development can only take place in individualized adaptive testing situations. This increases the likelihood that the child will be neither frustrated by item tasks that are too hard, nor bored by questions that are too easy.

Psychometrically, adaptive tests are significantly more efficient than a "one test form fits all" administrations since the reliability per unit of testing time is greater (Lord 1980). Adaptive testing also minimizes the potential for floor and ceiling effects, which can affect the measurement of gain in longitudinal studies. Floor effects occur when some children's ability level is below the minimum that is accurately measured by a test. This can prevent low-performing children from demonstrating their true gains in knowledge when they are retested. Similarly, ceiling effects result in failure to measure the gains in achievement of high-performing children whose abilities are beyond the most difficult test questions. In adaptive testing performance, the beginning of a testing session is used to direct the selection of later tasks of an appropriate difficulty level for each child. Adaptive testing relies on Item Response Theory (IRT) assumptions in order to place children who have taken different test forms on the same vertical score scale. More will be said about this when the psychometric characteristics of the direct cognitive measures are presented.

For these reasons, the Educational Testing Service (ETS) recommended that the ECLS–K use individually administered adaptive tests, and NCES accepted the recommendation. A review of commercially available tests indicated that there were no "off-the-shelf" tests that met the domain requirements and were both individually administered and adaptive.

2.2.2 Sources of the ECLS–K Frameworks

As stated earlier, the ECLS–K was charged with assessing cognitive skills that are both typically taught and developmentally important. Neither typicality nor importance was easily determined. Identifying typical curriculum objectives and their relative importance was difficult because of the decentralized control that characterizes the American education system. The difficulties were compounded for the ECLS–K, since curriculum is constantly evolving and the data collection was to start in 1998, two years after the design phase, and continue until 2004.

Fortunately, the ECLS–K was able to draw on the extensive work recently completed for the National Assessment of Educational Progress (NAEP) fourth-grade test specifications of 1992, 1994, and 1996. Some of the ECLS–K panel of consultants had been instrumental in developing the NAEP content and process frameworks for reading, mathematics, science, and social studies. The NAEP assessment goals are similar to those of the ECLS–K in that both projects aim to assess cognitive skills that schools typically emphasize. The NAEP frameworks were also very useful models since they begin at the fourth grade and thus define sets of skills and understandings that were appropriate for the later years of the ECLS–K. This overlap would allow for comparisons between the two studies and would potentially enrich what was learned from each of them. Since the properties of the ECLS–K vertical scales depend on linking items throughout the grades; item selection in the early grades should define a path to the fourth-grade NAEP specifications.

The NAEP 1992, 1994 and 1996 frameworks were based on both current curricula and recommendations for curriculum change that have strong professional backing among theorists and teacher associations. NAEP is interested in the recommendations because it is charged with assessing skills and knowledge that reflect "best practices," as well as those that are widely taught. In contrast, the ECLS–K examines the full range of practices rather than concentrating on best practices. Nonetheless, these recommendations represent reasonable predictions about the directions that schools and school systems in the United States are likely to take in the near future and were thus appropriate to the ECLS–K. With respect to current curricula, NAEP relied on advice from panels of curriculum specialists. In addition to often being directly involved in the development of curricula used in the schools, specialists often hold a wealth of local knowledge about current practices, which is not recorded in publications and thus not otherwise available.

Despite these strengths, the NAEP test specifications had some important limitations on their applicability to the ECLS–K. First, the NAEP specifications were developed for fourth grade and up, and thus may not be appropriate in some respects for the very early years in school. The NAEP fourth-grade reading assessment framework, for example, is based entirely on sentence- and passage-level reading comprehension, and these skills are well beyond the grasp of most kindergartners and first-graders. These kinds of disjunctures required the ECLS–K to modify some of the NAEP frameworks to better represent the early elementary years.

Secondly, the NAEP frameworks defined a number of different subscales within subject-matter domains, but test-length constraints forced the ECLS–K to define single proficiency scales for each subject domain. NAEP can measure multiple subscores within a content domain because it administers a large number of different item sets in a spiraled design to students at a given grade level. That design follows from NAEP's primary goal of measuring cognitive status at the aggregate level on a cross-sectional basis. In contrast, the ECLS–K attempts to attain relatively accurate longitudinal measurement (through adaptive test instrumentation and vertical scaling) at the individual level within a more focused cognitive domain.

For the grades in which the NAEP frameworks proved to be inappropriate, the ECLS–K relied primarily on advice from early elementary school educators and curriculum specialists to articulate more suitable test specifications. Their recommendations are described in the sections that follow on the specific subject-area tests.

With certain exceptions, most notably reading, the following proposed frameworks assume that the general specifications in each of the three content areas apply to all grades, but that the emphasis will change from grade to grade. These changes are reflected in the frameworks by changes in the percentages of the testing time that are allocated to measuring any given skill or cognitive process. This coherency of specifications across grades is consistent with the various sets of standards that were being published in the areas of mathematics, English language arts, social studies, and science.

It is important to bear in mind that the adaptive nature of the assessment is designed so that, for example, a first-grade student who does very well on the first-stage routing test in mathematics would receive a more difficult first-grade mathematics form that would include items from the second-grade specifications. Conversely a child who does very poorly on the same first-grade routing test would receive a relatively easy second-stage form that would include items from the kindergarten specifications. Children who perform at the grade average on the routing test would receive a second-stage form that most closely reflects the test specifications of their present grade. Note that the routing tests are always specific to a single subject area and affect the difficulty of the test taken only within that subject area. In other words, a child who does poorly on the mathematics routing test and takes a relatively easy mathematics form may do very well on the routing test for reading and thus take a relatively difficult reading test.

2.2.3 Item and Time Allocations

In addition to the conceptual framework identifying the various types of skills and knowledge tested in the ECLS–K, guidance was also needed on the relative emphases that the different outcomes should receive. The general rule that the ECLS–K used in determining allocations is that the compositions of the tests reflect typical curriculum emphases. Systematically collected evidence on typical curricular contents was not available in most subject areas, however, so the study relied mainly on an expert panel composed of curriculum specialists and people with extensive teaching and administrative experience in elementary schools. The overall testing time for each child was expected to consist of equal amounts of time for reading and mathematics, with a lesser amount of time allocated for the general knowledge test. Following the model of the NAEP 1996 mathematics framework, the ECLS–K chose to quantify relative emphases that should be devoted to each skill. It is important to keep in mind that some areas can be assessed more quickly than other areas (e.g., many vocabulary items can be administered in a short period of time, while passage comprehension items take longer to administer). Tables 2-1 to 2-4 present the test specifications for the ECLS–K cognitive battery from kindergarten to the fifth grade. The numbers in the cells are the target percentages *of testing time* for each content category; they are at best approximations since the item classifications are somewhat arbitrary. Particularly in the intermediate grades (e.g., 3 to 5), many items tap more than one area. For example, a mathematics problem may require skill in interpreting data as well as skill in understanding number concepts.

The ECLS–K tests include about 50 to 70 items per subject area test for each grade level. As noted earlier, there are some discrepancies between the time allocations and the number of items in each category, because some kinds of items usually take longer to administer than others. Reading comprehension items based on passages, for example, take longer than vocabulary items; mathematics items that require problem solving or computations take longer than pattern recognition items.

2.2.4 Mathematics Test Specifications

The mathematics test specifications shown in table 2-1 are primarily based on the *Mathematics Framework for the 1996 National Assessment of Educational Progress* (National Assessment Governing Board [NAGB] 1996a). The NAEP mathematics framework is itself largely based on the curriculum standards from the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics (NCTM 1989). The NCTM K-4 curriculum standards are listed in appendix A.

Two differences between the NCTM curriculum standards and the NAEP framework should be noted. One is that NAEP classified cognitive processes (conceptual understanding, procedural knowledge, and problem solving) as a separate dimension and cross-classified the cognitive processes with a subset of the NCTM content or strand classifications. ECLS–K addresses these cognitive processes within each content strand.

The content strands represented by the column categories in table 2-1 are defined as follows (these correspond closely to NAGB (1996a) definitions for most strands):

- *Number Sense, Properties, and Operations.* This refers to children's understanding of numbers (whole numbers, fractions, decimals, and integers), operations, and estimation, and their application to real-world situations. Children are expected to demonstrate an understanding of numerical relationships as expressed in ratios, proportions, and percentages. This strand also includes understanding properties of numbers and operations, ability to generalize from numerical patterns, and verifying results.

- *Measurement.* Measurement skills include choosing a measurement unit, comparing the unit to the measurement object, and reporting the results of a measurement task. It includes items assessing children's understanding of concepts of time, money, temperature, length, perimeter, area, mass, and weight.

- *Geometry and Spatial Sense.* Skills included in this content area extend from simple identification of geometric shapes to transformations and combinations of those shapes. The emphasis of the ECLS–K is on informal constructions rather than the traditional formal proofs that are usually taught in later grades.

- *Data Analysis, Statistics, and Probability.* This includes the skills of collecting, organizing, reading, and representing data. Children are asked to describe patterns in the data, or making inferences or drawing conclusions based on the data. Probability refers to making judgments about the likelihood of something occurring based on information collected on past occurrences of the event in question. Students answer questions about chance situations, such as the likelihood of selecting a marble of a particular color in a blind draw when the numbers of marbles of different colors are known.

- *Patterns, Algebra, and Functions.* Consistent with the NCTM kindergarten to fourth-grade curriculum standards, the ECLS–K framework groups pattern recognition together with algebra and functions. Patterns refer to the ability to recognize, create, explain, generalize, and extend patterns and sequences. In the kindergarten test, the items included in this category entirely consist of pattern recognition items. As one moves up to the subsequent grades, algebra and function items are added. Algebra refers to the techniques of identifying solutions to equations with one or more missing pieces or variables. This includes representing quantities and simple relationships among variables in graphical terms. It should be noted that while pattern recognition is relatively heavily emphasized in kindergarten and even first-grade classrooms, the proposed framework tends to de-emphasize the assessment allocation since it is not clear what to expect with reference to longitudinal trends in this skill area.

The time allocation targets listed in table 2-1 for the third, fourth, and fifth grades are close to the NAEP 1996 fourth-grade mathematics recommendations. NAEP recommends 40 percent of the fourth-grade items measure Number Sense, Properties, and Operations; 20 percent in Measurement; 15 percent in Geometry and Spatial Sense; 10 percent in Data Analysis, Statistics, and Probability; and 15 percent in Patterns, Algebra, and Functions. NAEP further recommends that at least half of the items in Number Sense, Properties, and Operations involve some aspect of estimation or mental mathematics.

The number sense, properties, and operations content strand represents the dominant emphasis of elementary school mathematics. The ECLS–K framework targets the development in this area through the fifth grade. There is a slight decrease in the assessment allocation after second grade from 50 percent in K-2 to 40 percent in the third to fifth grades, but this content strand is the largest in all grades included in the ECLS–K.

Table 2-1.— ECLS–K mathematics longitudinal test specifications, in percentages of testing time, for kindergarten through fifth grade

	Content strands[*]					
Mathematics processes	Number sense, properties, and operations	Measurement	Geometry and spatial sense	Data analysis, statistics and probability	Patterns, algebra, and functions	Totals
Kindergarten	50	15	5	10	20	100
First grade	50	14	10	10	16	100
Third grade	40	20	15	10	15	100
Fourth grade	40	20	15	10	15	100
Fifth grade	40	20	15	10	15	100

[*] The content strands are identical to those used in the "Mathematics Framework for the 1996 National Assessment of Educational Progress (NAEP)," (NAGB, 1996a). The content strand item targets for the third, fourth, and fifth grades match the NAEP fourth grade recommendations for the minimum number of "Number Sense" items, and the maximum numbers for the other strands. See the text for a discussion of the overlaps and disjunctions with the NCTM standards.

BEST COPY AVAILABLE

2.2.5 Reading Test Specifications

The ECLS–K reading specifications (table 2-2) were derived mainly from the Reading Framework for the 1992 and 1994 NAEP (NAGB 1994a). Literacy curriculum specialists were also consulted, and focus groups of kindergarten through second grade teachers were assembled to review the proposed framework and item pool.

The conceptual categories shown in table 2-2 are from the NAEP reading framework and the recommendations of the literacy curriculum specialists. The NAEP framework is defined in terms of four types of reading comprehension skills:

- *Initial understanding* requires readers to provide an initial impression or global understanding of what they have read. Identifying the main point of a passage and identifying the specific points that were drawn on by the reader to construct that main point would be included in this category.
- *Developing interpretation* requires readers to extend their initial impressions to develop a more complete understanding of what was read. It involves the linking of information across parts of the text, as well as focusing on specific information.
- *Personal reflection and response* requires readers to connect knowledge from the text with their own personal background knowledge. Personal background knowledge in this sense includes both reflective self-understanding, as well as the broad range of knowledge about people, events, and objects that children bring to the task of interpreting texts.
- *Demonstrating a critical stance* requires the reader to stand apart from the text and consider it objectively. This is includes questions asking about the adequacy of evidence used to make a point, or the consistency of someone's reasoning in taking a particular value stance. In kindergarten and first grade, some questions about unrealistic stories were asked to assess the child's notion of "real vs. imaginary." Such story types allow us to get information on critical skills as early as kindergarten.

Since the NAEP framework begins with fourth grade, it had to be modified to adequately accommodate the basic skills typically emphasized in the earliest grades. The ECLS–K thus added two additional skill categories to the NAEP framework: Basic Skills, which includes familiarity with print and recognition of letters and phonemes, and Vocabulary. However, the ECLS–K reading framework by fourth grade is very close to that of NAEP.

Notably absent from the ECLS–K reading framework is any place for writing skills. This absence is a reflection of practical constraints associated with cost of scoring and limited amount of testing time. It is also important to note that the ECLS–K asks teachers to provide information on each sampled child's writing abilities each year and on the kinds of activities they use in their classrooms to promote writing skills.

The time allocations shown in table 2-2 were developed by the ECLS–K advisors (NAEP provides little guidance on these decisions in the area of reading). The general approach followed by the ECLS–K in developing the reading assessment was to begin with relatively more emphasis on basic reading skills during the first years (kindergarten and first grade), decreasing as more emphasis is placed on measuring reading comprehension skills in the later years (fourth and fifth grade). The emphasis in the assessment of reading comprehension is on the inferential understanding of text or on developing interpretation. However, this does not mean that the basic reading skills of children in the third, fourth, and fifth grades will not be tested. With the adaptive nature of the test administration, children who do not perform well on their grade-specific routing test are assessed using a form with a lower level of difficulty. For example, a fourth-grader who does not perform well on the fourth-grade routing test is administered a form which would include relatively more basic skill items than would a child who had surpassed the basic level of achievement in reading.

The NAEP fourth-grade reading assessment framework distinguishes between reading for literary experience and reading for information. Consistent with NAEP, the ECLS–K roughly balances the number of items tied to fictional and informational texts.

Table 2-2.— ECLS–K reading longitudinal test specifications, in percentages of testing time, for kindergarten through fifth grade

Grade levels	Basic skills[1]	Vocabulary	Reading comprehension skills: Initial understanding[2]	Reading comprehension skills: Developing interpretation[3]	Reading comprehension skills: Personal reflection[4]	Reading comprehension skills: Critical stance[5]	Totals
Kindergarten	40	10	10	25	10	5	100
First grade	40	10	10	25	10	5	100
Third grade	15	10	15	30	15	15	100
Fourth grade	10	10	15	30	15	20	100
Fifth grade	10	10	15	30	15	20	100

NOTE: The column headings are identical to the NAEP 1994 Reading Framework categories, except ECLS-K added basic skills and vocabulary.

[1] Basic skills include familiarity with print and recognition of letters and phonemes.

[2] Initial understanding requires readers to provide an initial impression or global understanding of what they have read.

[3] Developing interpretation requires readers to extend their initial impressions to develop a more complete understanding of what was read.

[4] Personal reflection and response requires readers to connect knowledge from the text with their own personal background knowledge. The focus here is relating text to personal knowledge

[5] Demonstrating a critical stance requires the reader to stand apart from the text and consider it objectively.

BEST COPY AVAILABLE

2.2.6 General Knowledge: Science and Social Studies Test Specifications

The ECLS–K general knowledge test for kindergarten and first grade is approximately evenly divided between items that measure knowledge and skills in the natural sciences and social studies items. While these items may define a single "general knowledge" scale in the early elementary grades, the test specifications of science and social studies are separated because that allows researchers to identify better the kinds of knowledge and skills the ECLS–K is designed to measure. In later grades, only science is directly assessed in the ECLS–K.

2.2.6.1 Science

The test specifications for science were developed largely from recommendations of the ECLS–K advisory group. Similar to the 1996 NAEP Science Framework (NAGB 1996b), the ECLS–K science framework includes two broad classes of science competencies: Conceptual Understanding and Scientific Investigation.

- *Conceptual Understanding* refers to both the child's factual knowledge base and the conceptual accounts that children have developed for why things occur as they do. Consistent with current curriculum trends, the emphasis in the ECLS–K will be more on the adequacy of accounts than the grasp of discrete facts, particularly as the children move up in grade level.
- *Scientific Investigation* refers to children's abilities to formulate questions about the natural world, to go about trying to answer them on the basis of the tools available and the evidence collected, and to communicate their answers and how they obtained them.

The ECLS–K general knowledge test includes items drawn from the fields of earth, physical, and life science. These fields are defined as follows:

- *Earth and space science* is the study of the earth's composition, process, environments, and history, focusing on the solid earth and its interactions with air and water. The content to be assessed in earth science centers on objects (soil, minerals, rocks, fossils, rain, clouds, and the sun and moon), as well as processes and events that are relatively accessible or visible. Examples of processes are erosion and deposition and weather and climate; events include volcanic eruptions, earthquakes, and storms. Space science in the early elementary grades is usually concerned the relationships between earth and other bodies in space (e.g., patterns of night and day, the seasons of the year, and phases of the moon).
- *Physical science* includes matter and its transformations, energy and its transformations, and the motion of things.

- *Life science* is devoted to understanding and explaining the nature and diversity of life and living things. The major concepts to be assessed relate to interdependence, adaptation, ecology, and health and the human body.

In terms of subject-matter emphases in the elementary grades, the 1996 NAEP Science Framework, American Association for the Advancement of Science (AAAS 1995) and National Academy of Sciences (NAS 1995) recommend roughly equal emphasis on the three strands: earth, life, and physical science. Review of elementary text series (Harcourt Brace 1995, Ramsey 1986, Scott-Foresman 1994, and Silver Burdett & Ginn 1991) revealed that coverage of these topics is equally distributed. The ECLS–K advisors concurred with the recommendation of equal representation of the strands at each grade level, and the final item batteries reflect that balance. The ECLS–K science framework is shown in table 2-3.

Table 2-3.—ECLS–K science longitudinal test specifications, in percentages of testing time, for kindergarten through fifth grade

Grade level	Earth and space science	Physical science	Life science	Total
Kindergarten	33	33	33	100
First grade	33	33	33	100
Third grade	33	33	33	100
Fourth grade	33	33	33	100
Fifth grade	33	33	33	100

NOTE: The science expert panel on the ECLS–K developed the column categories and target allocations. The allocation of items at each grade level follows the 1996 NAEP guidelines that about half of the items within each of the science subdomains measure conceptual understanding and half measure scientific investigation.

BEST COPY AVAILABLE

2.2.6.2 Social Studies

The National Council for the Social Studies (1994) defines social studies as ". . . the integrated study of the social sciences and humanities to promote civic competence. Within the school program, social studies provides coordinated, systematic study drawing upon such disciplines as anthropology, archeology, economics, geography, history, law, philosophy, political science, psychology, religion, and sociology, as well as appropriate content from the humanities, mathematics, and natural sciences. The primary purpose of social studies is to help young people develop the ability to make informed and reasoned decisions for the public good as citizens of a culturally diverse, democratic society in an interdependent world."

The ECLS–K social studies framework is shown in table 2-4.

The column categories are simplifications of the early grade recommendations of the 1994 *Curriculum Standards of Social Studies* published by the National Council for the Social Studies (NCSS).

- *History* refers to knowledge of the ways people view themselves in and over time. (NCSS category "Time, Continuity, and Change".)
- *Government* refers to understandings of how people create and change structures of power, authority, and governance, as well as of the ideals, principles, and practices of citizenship in a democratic republic. (This includes items measuring the NCSS categories "Power, Authority, and Governance" and "Civic Ideals and Practices".)
- *Culture* includes knowledge about similarities and differences among groups, as well as about how individuals interact and understand themselves and others within a culture. (NCSS categories "Culture," "Individuals, Groups, and Institutions," and "Individual Development and Identity".)
- *Geography* refers to understanding of places, distances, and physical environments and how they shape and reflect people and their relations with others. (NCSS category "People, Places, and Environments".)
- *Economics* includes understandings of how people organize for the production, distribution, and consumption of goods and services. (NCSS category "Production, Distribution, and Consumption".)

Table 2-4.—ECLS–K social studies longitudinal test specifications, in percentages of testing time, for kindergarten through first grade

Grade level	History	Government	Culture	Geography	Economics	Total
Kindergarten						
Knowledge	8	8	40	16	8	80
Analysis and interpretation	2	2	10	4	2	20
Total	10	10	50	20	10	100
First grade						
Knowledge	7	7	35	14	7	70
Analysis and interpretation	3	3	15	6	3	30
Total	10	10	50	20	10	100
Total	10	10	50	20	10	100

History in kindergarten through first grade includes learning to distinguish between present and past. It is often centered in lessons tied to signal events and persons in American history and its larger cultural traditions, but can also include the history of ordinary families and groups.

Lessons about the government in the elementary curriculum can include concepts of the purposes of government; individual rights and responsibilities (often taught in relation to the children's families, peer groups, and school classes); and distinctions between local, state, and national government and their respective main officials.

The culture category in the ECLS–K kindergarten through first-grade tests includes a number of questions about everyday objects and their uses ("What do trains and planes have in common?") and social roles ("What does a fireman do?").

Geography in the early grades typically includes learning about where one lives in relation to the rest of the nation and the world, gaining familiarity with maps and the globe, and learning about different types of land and water and how people, plants, and animals have adapted to them (see also NAGB 1994b).

In the elementary grades, economics includes distinguishing between needs and wants, understanding rudiments of the division of labor (who does what and why there are so many different jobs), and the relationship of price to supply and demand.

The allocation of items to these different content areas is based on advice from curriculum specialists. The concepts and skills taught in kindergarten and first grade tend to group mainly in the Culture domain, with relatively little emphasis on the other content areas.

2.3 Indirect Cognitive Assessment: Academic Rating Scale

The academic rating scale (ARS) indirect cognitive measures were developed for the ECLS–K to measure teachers' evaluations of students' academic achievement in the three domains that are also directly assessed in the cognitive battery: language and literacy (reading), general knowledge (science and social studies), and mathematical thinking. The ARS was designed both to overlap and to augment the information gathered through the direct cognitive assessment battery. Although the direct and indirect instruments measure children's skills and behaviors within the same broad curricular domains with some intended overlap, several of the constructs they were designed to measure differ in significant ways. Most importantly, the ARS includes items designed to measure both the process and products of children's learning in school, whereas the direct cognitive battery assesses only the products of children's achievement. The scope of curricular content represented in the indirect measures is designed to be broader than the content represented on the direct cognitive measures. Because of practical constraints of testing time and format limitations, the direct cognitive battery was not able to assess writing skills or the strategies children use to solve problems.

Unlike the direct cognitive measures, which were designed to measure gain on a longitudinal vertical scale from kindergarten entry through the end of first grade, the ARS is targeted to a specific grade level. The questions range from explicitly objective items (e.g., "names all upper- and lower-case letters of the alphabet") to others with a more subjective element (e.g., "composes simple stories" or "uses a variety of strategies to solve mathematics problems"). Teachers evaluating the children's skills were instructed to rate each child compared to other children of the same age level.

The development of the indirect measures paralleled the development of the direct measures. A background review of the literature on the reliability and validity of teacher judgments of academic performance was conducted (see Meisels and Perry 1996). National and state standards as well as the literature on the predictive validity of early skills were examined to develop the item pool. The following criteria were used in creating and selecting items for the ARS:

- Skills, knowledge, and behaviors reflecting most recent state and national curriculum standards and guidelines;
- Variables identified in the literature as predictive of later achievement;
- Direct criterion-referenced items with high level of specificity that call for low levels of teacher inference;
- Skills, knowledge, and behaviors that are easily observable by teachers;
- Items broad enough to allow for diverse populations of students to be evaluated fairly;
- Some items that overlap with the content assessed through the direct cognitive battery;
- Some items that expand the skills tested by the direct cognitive battery—particularly those that assess process skills that would be difficult to assess given the time constraints;

- Literacy items that target listening, speaking, reading, and writing skills; and
- Items that reflect developmental change across time.

As listed here, among the criteria used in item construction was the ability to measure developmental growth over time. This was accomplished by including items that target the same skill, type of knowledge, or behavior across two or more assessment periods in the ECLS. These items were constructed to measure the same construct over time taking into account how skills, knowledge, and behaviors manifest themselves differently at various chronological and/or developmental periods. Although the measurement of many skills remains constant across grade levels (e.g., "understanding the conventions of print"), the item exemplars representing these skills increase in complexity as children progress through the grades. Increasing the complexity of exemplars over time is necessary in order to represent how constructs evidence themselves along a developmental continuum.

Teachers were to rate each child's skills, knowledge, and behaviors on a scale from "Not Yet" to "Proficient" (see table 2-5). If a skill, knowledge, or behavior had not yet been introduced into the classroom, the teacher coded that item as N/A (not applicable). The differences between the direct and indirect cognitive assessments and the scores available are described here. For a discussion of the content areas of the ARS, see chapter 2, section 2.4.1 of the ECLS–K user manuals.

Table 2-5.—Academic rating scale response scale

	Rating	Description
1	Not yet	Child *has not yet* demonstrated skill, knowledge, or behavior.
2	Beginning	Child is *just beginning* to demonstrate skill, knowledge, or behavior but does so very inconsistently.
3	In progress	Child demonstrates skill, knowledge, or behavior *with some regularity* but varies in level of competence.
4	Intermediate	Child demonstrates skill, knowledge, or behavior *with increasing regularity and average competence* but is not completely proficient.
5	Proficient	Child demonstrates skill, knowledge, or behavior *competently and consistently.*
	N/A	Not applicable: Skill, knowledge, or behavior has *not been introduced* in classroom setting.

Kindergarten and grade-one teachers from both public and private schools and content experts familiar with the early grades reviewed the items and made recommendations. Items were then piloted and later field-tested in order to gather statistical evidence of the appropriateness of the items for carrying out the overall assessment goals. The pilot testing indicated that the difficulty of the items needed to be increased in order to capture the range of abilities represented in the early grades and to avoid a serious ceiling problem. The items were revised and the difficulty of the criteria in the exemplars increased before field testing. The items were field tested in the spring of 1997 during the field test of the direct cognitive assessments. Final items were chosen consistent with the item statistics and representativeness of the content.

2.4 Social Rating Scales: Teacher and Parent

The social rating scale (SRS) is an adaptation of the Social Skills Rating System (Gresham & Elliott 1990). Both the teacher and parent use a frequency scale (see table 2-6) to report on how often the student demonstrates the social skill or behavior described. Factor analyses (both exploratory analyses and confirmatory factor analyses using LISREL) were used to confirm the scales. See chapter 2, section 2.3 and 2.4 of the ECLS–K user manuals for additional information on the parent and teacher SRS instruments.

Table 2-6.—Social rating scale response scale

Rating		Description
1	Never	Student never exhibits this behavior.
2	Sometimes	Student exhibits this behavior occasionally or sometimes.
3	Often	Student exhibits this behavior regularly but not all the time.
4	Very often	Student exhibits this behavior most of the time.
N/O	No opportunity	No opportunity to observe this behavior.

The items on the parent SRS were to be administered as part of a telephone or in-person survey. (See chapter 2, section 2.3 in the ECLS–K user manuals for a more detailed description of the parent scales.) The factors on the parent SRS are similar to the teacher SRS; however, the items in the parent SRS are adapted to the home environment and, thus, are not the same as the teacher items. It is also important to keep in mind that parents and teachers observe the children in very different environments.

2.5 Psychomotor Assessment

The psychomotor assessment is an adaptation of the motor scale of the Early Screening Inventory-Revised (Meisels, Marsden, Wiske, & Henderson 1997). The total score includes two scales, one measuring fine motor skills (eye-hand coordination) and the other measuring gross motor skills (balance and motor planning). The fine motor skills score is the sum of the points for seven tasks: build a gate, draw a person, and copy five simple figures. Children could receive up to two points for each of the first two tasks and one point for each of the figures. Gross motor skills consisted of balancing, hopping, skipping, and walking backward—children could receive up to two points for each skill. Confirmatory factor analysis during the ECLS–K design phase (using LISREL) confirmed the two scales.

2.6 Oral Language Development Scale

An objective of the ECLS–K was to include language minority children in all survey activities to the extent permitted by their English proficiency. A panel convened by the ECLS–K design team recommended that an English proficiency test, rather than recommendations of parents or teachers, be used to evaluate language minority children's ability to participate in the direct cognitive testing. Since states and school districts vary in the criteria they use to identify children's English proficiency, a single standard consistently applied to all children in the sample was suggested.

Staff at the American Institutes for Research (Montgomery 1997) carried out an investigation to identify an appropriate English language proficiency measure. The selected measure needed to be relatively short, easy to administer, and easy to score. In addition, it should have known psychometric properties, including predictive validity and face validity among experts in the field. On the basis of a literature search, advice from experts in language minority assessment issues, and information from departments of education in the four states with the largest percentages of language minority individuals, five tests were identified as possible candidates.

The consultants recommended the PreLAS 2000 (Duncan & DeAvila 1998), the pre-kindergarten/kindergarten/first grade version of CTB/McGraw-Hill's Language Assessment Scales (LAS), for several reasons, including the following:

- Widespread use and acceptance for the age group,
- Content matching the ECLS–K requirements, and
- Similarity to the ECLS–K cognitive battery in format and administration procedures.

The Pre-LAS 2000 consists of six scales, measuring both receptive and productive language. Edward De Avila, a co-author of the PreLAS 2000, consulted with ECLS–K project staff in selecting three of the six scales of PreLAS 2000 Form C to serve as the English screening test for the ECLS–K. A Spanish version of the Oral Language Development Scale (OLDS) was created as well, consisting of the equivalent three subtests from the PreLAS Espanol (Duncan & De Avila 1986). The Spanish version measured the same constructs measured by the English version, using the same activities but with different stories and stimulus pictures. The subtests making up the English and Spanish OLDS for the ECLS–K were as follows:

- "Simon Says" ("Tio Simon") measured listening comprehension of simple directives in English/Spanish (i.e., asking a child to do things such as touch ear, pick up paper, or knock on table).
- "Art Show" ("La Casita") was a picture vocabulary assessment where children were asked to name pictures they were shown. The Art Show served as an assessment of a child's oral vocabulary.
- "Let's Tell Stories" ("Contando Historias") was used to obtain a sample of a child's natural speech by asking the child to retell a story read by the assessor. The child was read two different stories (selected at random from three possibilities) and asked to retell it in his or her own words using pictures as prompts. Scores were based on the complexity of the child's sentence structure and vocabulary in his or her retelling of the story.

The first two subtests consisted of ten items each, scored one point per item. The story subtest was scored 0 to 5 points for each story and weighted at four times the Simon Says and Art Show items, for a total of 60 possible points for the three subtests selected for the OLDS. Dr. De Avila recommended requiring a score of at least 37 out of 60 as the level at which children understood English well enough to receive the direct child assessment in English. This cutting score was based on results of a national norming sample for PreLAS, extrapolated to the three selected subtests. Children who scored 36 or below, and whose native language was not Spanish, were excluded from the direct cognitive assessment. Spanish speakers who scored 36 or below were administered the Spanish form of the OLDS as a measure of their proficiency in Spanish. They then proceeded to take Spanish language versions of the ECLS–K mathematics and psychomotor assessments.

Field supervisors either checked school records to determine children's home language or, if records were not available, requested this information directly from children's teachers. The OLDS was given to those children who had a non-English language background. Children who did not achieve the cutting score during one round of data collection were screened again at the next round of testing to determine whether their English language skills had progressed to the point where they could be assessed in English. Once a child reached the target score of 37 or above, he or she was not rescreened in subsequent rounds but proceeded directly to the cognitive assessments.

3. DEVELOPMENT OF THE TWO-STAGE DIRECT TEST FORMS

This chapter describes the development of the item pool, the procedures used in the field test, the subsequent item analysis, and building of the two-stage forms.

3.1 Development of the Item Pool

Given the blueprints from the test specifications, the contractor assembled item writers from Educational Testing Service (ETS), elementary school curriculum specialists, and kindergarten, first and second grade teachers to develop items. Pools of slightly over 200 items in each of the three content domains were developed. Some of the items were borrowed or adapted, with permission, from published tests, including the Peabody Individual Achievement Test-Revised (PIAT-R), Peabody Picture Vocabulary Test-Revised (PPVT-R), the Primary Test of Cognitive Skills (PTCS), the Test of Early Reading Ability (TERA-2), The Test of Early Mathematics Ability (TEMA-2), and the Woodcock-Johnson Tests of Achievement-Revised (WJ-R).

The pools of items were reviewed for appropriateness of content and difficulty, and for relevance to the test framework. In addition items were reviewed for sensitivity issues related to minority concerns. Items that passed these content, construct, and sensitivity screenings were assembled into field test booklets.

3.2 Field Testing and Item Analysis

The field test was set up to shed light on at least four issues in addition to gathering the necessary psychometric data. One issue was whether it was possible to take children who were not yet reading and had limited numeracy skills (most fall-kindergartners) and put them on the same vertical scale as children who were reading (e.g., many spring-first-graders). The second issue was related to the attention span of the fall-kindergartners and whether they could complete the battery in one sitting without showing signs of distress. The third issue pertained to whether the individualized two-stage testing procedure with "on-time" scoring of the routing test would prove to be operationally feasible.

Finally, the items selected for the reading domain were to be validated by comparison with an established assessment instrument.

Approximately 100 to 120 items were field tested in each of the three cognitive domains, reading, mathematics, and general knowledge. Within each domain, the items were divided into two approximately parallel blocks, "A" and "B". The blocks were spiraled within seven test booklets; that is, each block of items appeared once in the first position in a booklet, once in the second position, and once as the last block, so that influences on performance, due to either fatigue or practice, would be minimized. Also, each block of items was paired with each other block in one booklet, so that correlations within and across content domains could be computed. A block of psychomotor items was also prepared, which included both fine motor and gross motor tasks. Each child received three blocks of field test items. See the field test report (Ingels et al. 1997) for additional information on the test design.

In fall 1996, one of the seven field test booklets was administered to each of 1,500 kindergarten children, resulting in about 600 observations on each test item. These same children were followed up in the spring of 1997, thereby providing some longitudinal estimates of growth from fall- to spring-kindergarten. A sample of approximately 1,500 first graders was field tested in spring 1997 as well using the same set of items as for the kindergartners. A subset also received the reading section of the Kaufman Test of Educational Achievement (KTEA) as a check on the construct validity of the Early Childhood Longitudinal Study Kindergarten Class of 1998–99 reading items. The original study design did not call for testing in the fall of first grade (a subsample was later specified) so first graders were not included in the fall 1996 field test.

Classical item statistics as well as Item Response Theory (IRT) parameters (Lord 1980) were estimated. The IRT parameters were based on the three parameter model with a parameter for guessing, a parameter for difficulty, and a slope parameter. Marginal maximum likelihood estimation procedures (Mislevy & Bock 1982, Muraki & Bock 1991) were used to estimate the item parameters.

Item trace plots were inspected for indications of lack of fit. The item trace plots identified the residuals by their grade membership. Thus, items could be identified that fit across all three time periods (fall- and spring-kindergarten, and spring-first grade) or demonstrated a good fit for a subset of the three groups. For example, a subset of items might have demonstrated good fits for spring-first grade but not for the kindergarten data. A relatively small percentage of items (about 10 to 15 percent) exhibited overall lack of fit. These were removed from consideration for the kindergarten to first-grade battery. For some of the poorer fitting items, a distracter analysis indicated that one of the incorrect response options was drawing the higher scoring individuals leading to a zero or negative biserial and/or flat or negative "a" parameters. In some cases modifications to the distracters were made, and the item was kept in the pool. Attempts to modify and retain items were particularly important for items that represented one of the more difficult-to-fill cells in the framework classifications.

3.2.1 Differential Item Functioning Analysis

Cognitive test items were checked for Differential Item Functioning (DIF) for males compared with females and for Black and Hispanic students compared with White students. It is not necessarily expected that different subgroups of students will have the same average performance on a set of items. But when students from different groups are matched on overall ability, performance on each test item should be about the same. There should be no relative advantage or disadvantage based on the student's gender or racial/ethnic group.

The DIF procedure (Holland & Thayer 1986) is designed to detect possible differential functioning for subgroups by comparing performance for a focal group (e.g., females or Black students) with a matched reference group (e.g., males or White children). DIF refers to the identification of individual items on which members of some population subgroups (the focal groups) perform particularly poorly in comparison to a reference group that is matched in terms of performance on the total pool of items. Items are classified as "A," "B," or "C" depending on the statistical significance of subgroup differences, as well as effect sizes. Items identified as having "C" level DIF have detectable differences that are both sizeable and statistically significant.

A finding of differential functioning, however, does not automatically mean that the item is flawed. A judgment that these items are inappropriate for one or more subgroups requires not only the statistical measure of DIF but also a determination that the difference in performance is not related to the construct being measured. It simply means that it is differentially easier or more difficult for some subgroup (focal group) when compared with a reference group. In other words, different population subgroups may have differential exposure or skill in solving test items relating to a topic that is to be measured. If so, the finding of differential performance may be an important and valid measure of the targeted skill. Items that demonstrate differential functioning favoring the reference group were reviewed for inappropriate content by a standing committee on test fairness at ETS, consisting of members from both majority and minority groups. Items that were judged to have content or presentation that might be problematic for a particular focal group were dropped from the item pool. However, items that had DIF that was judged to be a result of possible differential skills in some area of the test framework, and not due to subgroup membership, were retained. A more complete discussion of DIF methodology can be found in chapter 4.

3.2.2 Field Test Conclusions

With respect to the first issue, scalability, the IRT goodness-of-fit results were sufficiently good to suggest that the issue of building a vertical scale that could span prereading to reading was virtually a moot point. The second issue, the child maintaining his or her attention span throughout the testing situation without undue stress, also seemed to have a favorable resolution. The majority of the children enjoyed the testing situation and welcomed the individual attention of the test administrator. The operational issue concerning the scoring of a first stage and directing the child to a second stage did not encounter any serious problems when tested out during a pilot of the computer-assisted administration of the battery.

An additional issue, evaluating the construct validity of the reading test, was accomplished by including the Kaufman Test of Educational Achievement (KTEA) reading test in one of the field test spiral blocks. Evidence for the construct validity of the ECLS–K reading item pool was supported by the fact that it correlated in the mid- to upper-eighties with the KTEA.

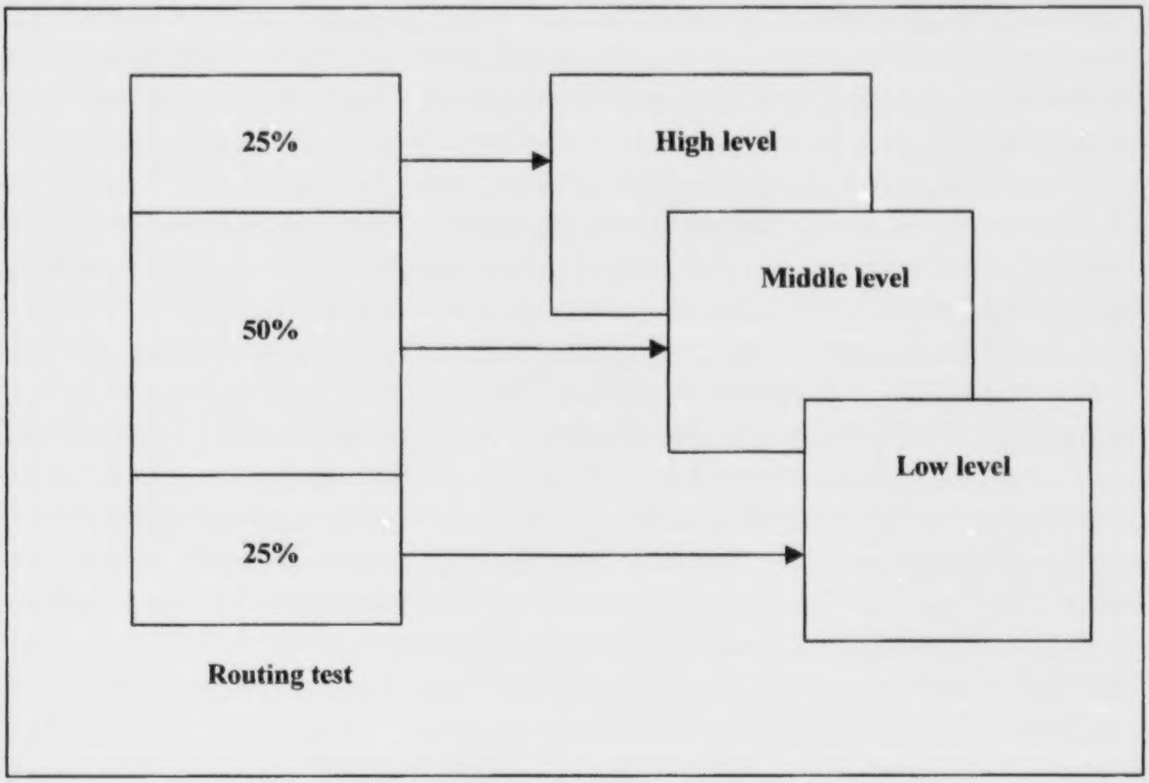
Field-tested items were candidates for final test forms if they had acceptable item analysis statistics and IRT parameters, had no DIF problems related to subgroup membership, and showed some increase in percent correct between fall-kindergarten and spring-first-grade.

3.3 Assembly of the Final Adaptive Forms

3.3.1 Two-Stage Testing Procedure

Figure 3.1 presents the general scheme for the two-stage testing procedure: a routing test and three second-stage forms. This figure illustrates the anticipated percentages of children in spring-kindergarten who would be routed to each of the second-stage forms. This scheme was followed in both reading and mathematics, while general knowledge had only two levels in the second stage. Figure 3-1 shows overlaps between the second-stage forms. This overlap serves two purposes. First, it insures a minimum of floor or ceiling effects even if a child happened to be assigned the wrong second-stage form. Secondly, it provides additional linking items to help anchor the vertical equating.

Figure 3-1.—Two-stage adaptive design



The contractors had used a similar two-stage adaptive test in the National Education Longitudinal Study:88 (NELS:88) (Rock et al. 1995), but that procedure was not "on-time" adaptive. NELS:88, which surveyed students in grades 8, 10, and 12, used group administration. In the grade 10 and 12 waves, a student was assigned to one of two reading forms varying in difficulty, and one of three mathematics forms, depending on how that same student performed on his or her previous testing two years earlier. Thus, the score based on the previous administration served as the routing test for the selection of the test form on the succeeding test administration. Since the ECLS–K is individually administered, a determination of the child's routing test score can be determined immediately, and he or she can be assigned the appropriate second-stage form immediately. It was reasoned that this "on-time" two-stage adaptive approach was particularly important in assessing growth in the early years since higher growth rates are expected on average in younger children, and one can also expect considerable variability in the individual growth rates. To capture both the extent and variability of growth, "on-time" adaptive testing seemed appropriate.

All children began the **reading** test with the same 20-item routing section. Depending on the child's score on the first stage or routing test, the child was assigned one of the three second-stage forms. The second-stage forms consisted of an easy form of 18 items, a middle-difficulty form containing 29 items, and a hard form, also with 29 items. Thus, a child would be administered 38 items in reading if he or she took the lower form and 49 items if his or her reading skills were sufficiently advanced for the middle or high form to be selected. Simulations from the field test suggested that 40 to 50 items per child would generate a target reliability of .90. Tests were discontinued at predetermined points if the child was struggling with the material or showing any distress.

The easy second-stage reading form had six items that overlapped with the middle-level form and seven that overlapped with both the middle- and high-level forms, plus five unique items. The middle-level second-stage form had nine unique items. The high second-stage form had 18 unique items. The common routing test and the item overlap between adjacent forms helped to insure that there would be sufficient numbers of good linking items to guarantee the stability of the vertical scale. The sharing of common items at the boundaries of the second-stage forms also minimizes measurement errors for those individuals who, for whatever reason, received second-stage forms that were not the ideal match for their achievement levels.

In designing a two-stage test, there are a number of decisions to be made. First, what proportion of the items to be administered should be allocated to the first stage? Secondly, having decided on the relative proportions, how does one set the cutting score on the first stage? The question of allocation of items to the first and second stages is often constrained by practical considerations. The first stage has to have items that can be quickly and unambiguously scored by the test administrator. This means extended free-response items would generally not be appropriate for the first stage. Long reading tasks also would not be an efficient use of time in the first stage since everyone takes the first stage, both readers and nonreaders. Long passages would provide more efficient measurement and use of time if placed in the middle- or high-level second-stage form where more children can read and do such tasks relatively fast. Given these practical considerations the first stage of the reading task consisted of quickly administered items spread over a broad range of difficulty.

Once the items for the routing test were selected, the items were then selected for the second-stage forms to fill in the gaps in difficulty not covered in the relatively broadband routing test. The relatively greater numbers of items in the middle and high form primarily reflected the fact that children going to these forms tended to spend less time per item than did the less skilled readers. Another objective in selecting more items for the higher-level forms was having enough difficult items to avoid ceiling effects.

Simulations based on the IRT ability estimates and item parameters solved the problem of defining the cutting scores on the first stage. Using the distributions of ability scores from the kindergarten and first-grade field test samples, simulations were carried out resulting in estimated frequency distributions of routing test scores. Cutting scores within these distributions were determined such that approximately 75 percent of fall-kindergarten children would be routed to the low-level second-stage form, and 75 percent of spring-first-graders would receive the most difficult form. Table 3-1 shows the cutting scores for routing and the percentage of students anticipated to take each second-stage form in the full-scale rounds of testing (not including the fall-first-grade subsample). Sections on samples and operating characteristics in chapter 5 (sections 5.2 through 5.4) present the actual percentages achieved in the operational rounds of testing. The success of the two-stage test design in achieving its goals is discussed there as well.

Table 3-1.—Routing cutting scores and anticipated second-stage percentages

Category	Low form	Middle form	High form
Reading			
Routing Test # Right	0-8	9-13	14-20
Planned % Fall-K	75%	20%	5%
Planned % Spring-K	25%	50%	25%
Planned % Grade 1	5%	20%	75%
Mathematics			
Routing Test # Right*	0-8	9-11	12-18
Planned % Fall-K	75%	20%	5%
Planned % Spring-K	25%	50%	25%
Planned % Grade 1	5%	20%	75%
General Knowledge			
Routing Test # Right	0-6	---	7-12
Planned % Fall-K	75%	---	25%
Planned % Spring-K	50%	---	50%
Planned % Grade 1	25%	---	75%

* Routing counts for the mathematics test included two practice items.

Second-stage items whose difficulty levels matched the target range of abilities were selected for each form. Additional easier and harder items were added to each form for the purposes of stabilizing the scale and avoiding floor and ceiling effects, as described earlier.

After the spring-kindergarten data had been collected and analyzed, the ability levels of the national sample were found to be somewhat higher than had been found in the field test. At that time, a supplementary set of 20 more difficult items was added to the high-level form for rounds 3 and 4.

The procedures for the assembly and identification of the cutting scores for the **mathematics** measure followed the same format as that of the reading test. The mathematics test had a 17-item routing test. Those assigned to the lower form received 18 more items, six of which were unique to the low form, and the rest were common to the middle-level or to both the middle- and high-level forms. Those children assigned to the middle-level form received 23 more items, five of which were unique to the middle form. Similarly those taking the high-level form received up to 31 more items, 18 of which were unique to the high form. The rationale for giving more items to the middle- and high-scoring children was the same as that given in the case of the reading measure.

The **general knowledge** test covered a less homogeneous content domain than did the reading and mathematics tests. The test specifications covered two domains: science and social studies, although the typical kindergarten curriculum does not include teaching a formal body of knowledge in these areas. At least in the first two rounds, most of the child's knowledge in these two areas may be largely the result of his or her family background, home educational environment, and preschool experiences. The title of general knowledge seems appropriate here, especially when the IRT model builds on the common factor underlying both domains.

The rationale and procedures used in reading and mathematics item selection for the routing and second-stage tests were also implemented here. However, because of the greater heterogeneity of content and less potential for school-related growth, it was decided to design only two second-stage forms. With only two forms, one could more easily balance the number of items from each of the two domains in each of the second-stage forms. The final routing test consisted of 12 items, with 25 items in the second-stage low-level form and 29 items in the second-stage high-level form. Ten items in the low-level form were unique to that form, while the high-level form had 14 unique items.

3.3.2 Criterion-Referenced Item Clusters

As indicated earlier, the ECLS–K was committed to reporting criterion-referenced scores as well as normative scores. Clusters of items provide a more reliable test of mastery or proficiency than do single-marker items because of the possibility of guessing. It is very unlikely that a child who has not mastered a skill defined by a cluster of marker items would be able to guess the correct answers to a majority of items in the cluster.

In consultation with curriculum specialists, five clusters of four items each were identified that marked agreed-on learning milestones in reading and mathematics. The five proficiency levels within each content area are assumed to follow a Guttman (1954) scale, that is, a child who passes a particular skill level (defined as any three of four items in the cluster correct) is expected to have mastered all the lower levels. A very small percentage of students in kindergarten and first grade had response patterns on the clusters that did not follow a Guttman scale. Overall, including all four rounds of data collection, only about six percent of the reading and five percent of the mathematics response patterns did not fit the hierarchical model.

Five clusters of items were selected that marked stages in going from prereading to reading. These item clusters of four items reflected skills that are typically taught in an ordered sequence. Items within a cluster had similar difficulties and shared similar skills. These item clusters formed a hierarchical structure in the Piagetian sense in that the teaching sequence implied that one had to master the lower levels in the sequence before one could learn the material at the next higher level. This theoretical and practical hierarchy was reflected in the ascending difficulties of the clusters of marker items. The five four-item clusters identified in the reading test were as follows:

- Level 1. Letter recognition: identifying upper and lower case letters by name.
- Level 2. Beginning sounds: associating letters with sounds at the beginning of words.
- Level 3. Ending sounds: associating letters with sounds at the end of words.
- Level 4. Sight words: recognizing common words by sight.
- Level 5. Comprehension of words in context: selecting the best word to complete a sentence.

An additional reading level "0" was hypothesized that would precede level 1 in the hierarchy above, and that consisted of three items targeting familiarity with conventions of print. However, this cluster did not fit the hierarchical model when the test responses were examined. As a result, separate "conventions of print" number-right scores were computed but were not considered to be part of the set of five hierarchical reading proficiencies.

A child was deemed proficient at any one level if he or she passed any three out of four items. An additional single item was then constructed for each of the five proficiency levels. A child was given a "1" on these supplemental items if he or she got any three out of four correct on each set of four items that marked the five proficiency levels; otherwise the score was zero. The creation of these "super items" and the subsequent estimation of their IRT parameters located the five proficiency levels on the reading score scale. This parameter estimation allows one also to estimate a continuous measure of the child's probability of being proficient at each of the five levels using the child's IRT ability estimate score and the parameters for each of the "super items."

Five clusters of four items were identified to mark milestones on the growth curve in mathematics. The five criterion referenced levels were as follows:

- Level 1. Number and shape: identifying some one-digit numerals, recognizing geometric shapes, and one-to-one counting of up to ten objects.

- Level 2. Relative size: reading all single-digit numerals, counting beyond ten, recognizing a sequence of patterns, and using nonstandard units of length to compare objects.
- Level 3. Ordinary number sequence: reading two-digit numerals, recognizing the next number in a sequence, identifying the ordinal position of an object, and solving a simple word problem.
- Level 4. Addition/subtraction: solving simple addition and subtraction problems.
- Level 5. Multiplication/division: solving simple multiplication and division problems and recognizing more complex number patterns.

The items within the mathematics clusters were somewhat more heterogeneous than was the case for reading, reflecting greater differences in the order of presentation of topics within the mathematics curriculum.

No criterion-referenced levels were postulated for the general knowledge test. Because of the two content domains and the diversity of curriculum in these areas, it would be difficult to argue for proficiency levels that would follow a hierarchical model *and* have logical interpretations.

3.3.3 English Fluency and Spanish Mathematics Test

The ECLS–K mathematics assessment was translated into Spanish and back translated by native Spanish speakers. The two versions were adjudicated and a penultimate version of the mathematics assessment was prepared. The Spanish mathematics assessment was sent to two expert reviewers, mathematicians who were native Spanish speakers, recommended by Richard Duran, a member of the ECLS–K Technical Review Panel, and Ed DeAvila, the developer of the LAS and PreLAS language screening tests. Comments from the expert reviewers were incorporated into the final version of the Spanish mathematics assessment. No such translation was developed for the other subject areas. The vocabulary, sight words, and rhyming items in the reading test could not be expected to have the same level of difficulty in a translated version. Nor would the general knowledge test, which is heavily culture-dependent, yield comparable measurement for children who are English-language-learners. The mathematics test, however, was judged to be much less dependent on language and culture, and was translated into Spanish. Translations of the assessments into languages other than Spanish were considered, but small sample sizes made this idea impractical.

Children who were not native speakers of English were given a screening test to determine whether their English language skills were sufficiently advanced to participate in the test battery. Children whose English skills had not permitted administration of the ECLS–K battery English language tests in the first round were rescreened in later rounds to ascertain whether their English fluency had reached the required level. Those who failed the English Oral Language Development Scale (OLDS) and were Spanish-speakers were then tested for fluency in Spanish and administered a Spanish translation of the mathematics test.

See chapter 5 for information on the performance of the Spanish mathematics assessment in the kindergarten and first-grade rounds of data collection.

4. ITEM RESPONSE THEORY SCALING FOR LONGITUDINAL MEASUREMENT AND EQUATING TO EARLIER ROUNDS

Measuring the extent of cognitive gains at both the group and individual level requires that the various kindergarten and first-grade forms must be calibrated on the same scale. The most appropriate way of doing this is to use Item Response Theory (IRT). To successfully carry out such a calibration, the sets of test items should be relatively unifactorial within a subject area (reading, mathematics, or general knowledge), with the same dominant factor underlying all test forms. This implies that there should be a common set of anchor items across adjacent forms and that most, but not necessarily all, content areas be represented in all grade forms. Increments in difficulty demanded in ascending grade forms (kindergarten to fifth grade) can be accomplished by (1) increasing the problem-solving demands within the same content areas and (2) including content in the later forms (in particular fourth and fifth grade) that tap materials normally found in the advanced course sequence but build on skills learned earlier in the sequence.

As indicated earlier, IRT (Lord 1980) was used in calibrating the various forms within each content area. A brief background on IRT follows with additional information on the Bayesian approach taken here.

4.1 Overview of Item Response Theory

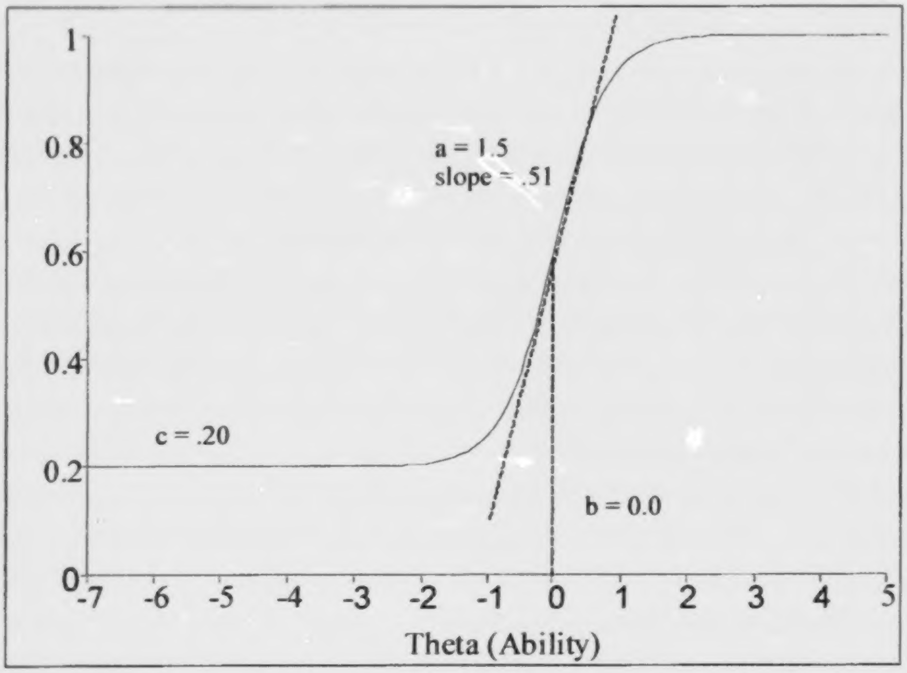
The underlying assumption of IRT is that a test taker's probability of answering an item correctly is a function of his or her ability level for the construct being measured and of one or more characteristics of the test item itself. The three-parameter IRT logistic model uses the pattern of right, wrong, and omitted responses to the items administered in a test form and the difficulty, discriminating ability, and "guess-ability" of each item, to place each test taker at a particular point, θ (theta), on a continuous ability scale. Figure 4-1 shows a graph of the logistic function for a hypothetical test item. The horizontal axis represents the ability scale, theta. The point on the vertical probability axis corresponding to the height of the curve at a given value of theta is the estimated probability that a person of that ability level will answer the test item correctly. The shape of the curve is given by the following equation describing the probability of a correct answer on item *i* as:

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-1.702 * a_i(\theta - b_i)}} \qquad (4.1)$$

where θ = ability of the test taker
a_i = discrimination of item i, or how well the item distinguishes between ability levels at a particular point
b_i = difficulty of item i
c_i = "guessability" of item i

The "c" parameter represents the probability that a test taker with very low ability will answer the item correctly. In figure 4-1, about 20 percent of test takers with a very low level of mastery of the test material guessed the correct answer to the question. The "c" parameter will not necessarily be equal to 1/(# options) (e.g., .25 for a four-choice item). Some response options may, for unknown reasons, be more attractive than random guessing, while others may be less likely to be chosen.

Figure 4-1.—Probability of correct answer



The IRT "b" parameters correspond to the difficulty of the items, represented by the horizontal axis in the ability metric. In figure 4-1, b = 0.0 means that test takers with $\theta = 0.0$ have a probability of getting the answer correct that is equal to halfway between the guessing parameter and 1. In this example, 60 percent of people at this ability level answered the question correctly. The "b" parameter also corresponds to the point of inflection of the logistic function. This point occurs farther to the right for more difficult items and farther to the left for easier ones. Figure 4-2 is a graph of the logistic functions for seven different test items, all with the same "a" and "c" parameters and with difficulties ranging from b = -1.5 to b = 1.5. For each of these hypothetical questions, 60 percent of test takers whose ability level matches the difficulty of the item are likely to answer correctly. Fewer than 60 percent will answer correctly at values of theta (ability) that are less than b, and more than 60 percent at $\theta > b$.

The discrimination parameter, "a", has perhaps the least intuitive interpretation of all. It is proportional to the slope of the logistic function at the point of inflection. Items with a steep slope are said to discriminate well. In other words, they do a good job of discriminating, or separating, people whose ability level is below the calibrated difficulty of the item (who are likely to get it right at only about the guessing rate) from those of ability higher than the item "b", who are nearly certain to answer correctly. By contrast, an item with a relatively flat slope is of little use in determining whether a person's correct placement along the continuum of ability is above or below the difficulty of the item. This idea is illustrated by figure 4-3, representing the logistic functions for two test items having the same difficulty and guessing parameters but different discrimination. The test item with the steeper slope (a = 2.0) provides useful information with respect to whether the test taker's ability level is above or below the difficulty level, 1.0, of the item: if the answer to this item was incorrect, the person very likely has an ability below 1.0; if the answer is correct, the test taker probably has a θ greater than 1.0, or guessed successfully. A series of many such highly discriminating items, with a range of difficulty levels (b parameters) such as those shown in figure 4-2, will do a good job in narrowing the choice of probable ability level. Conversely, the flatter curve in figure 4-3 represents a test item with a low discrimination parameter (a = .3). There is little difference in proportion of correct answers for test takers several points apart on the range of ability. So knowing whether a person's response to such an item is correct or not contributes relatively little to pinpointing his or her correct location on the horizontal ability axis.

With respect to interpreting the item parameters, "a" parameters (the discrimination parameter) should each be over .50; "a" parameters in the neighborhood of 1.0 or above are considered very good. As described earlier, the "a" parameter indicates the usefulness of the item in discriminating

Figure 4-2.—Items with different difficulty (b)

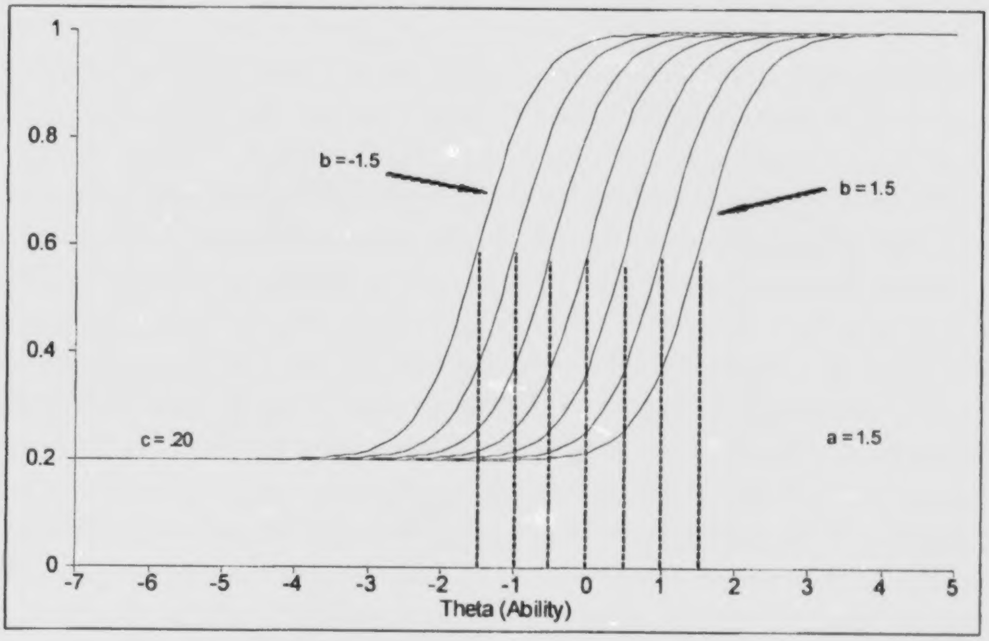
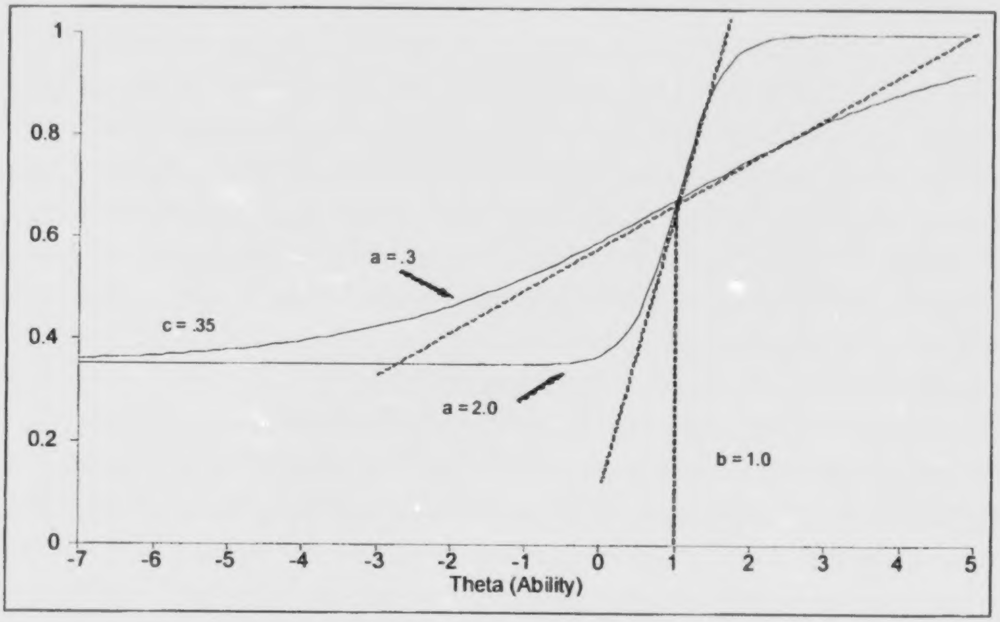


Figure 4-3.—Items with different discrimination (a)



between points on the ability scale. The "b" parameter, item difficulty, should span the range of abilities being measured. Item difficulties should be concentrated in the range of abilities that contains most of the test takers. Test items provide the most information when their difficulty is close to the ability level of the examinees. Items that are too easy or too difficult for most of the test takers are of little use in discriminating among them. Ideally the "c" parameter (the probability of a low ability person guessing correctly) tends to be less than .25 for four choice items, but may vary with difficulty, and of course the number of options. Open-ended items typically have a "c" parameter that is close to zero. In general, the ECLS–K item parameters meet these standards.

Once a pool of test items exists whose parameters have been calibrated on the same scale as the test takers' ability estimates, a person's probability of a correct answer for each item in the pool can be computed, even for items that may not have been administered to that individual. The IRT-estimated number correct for any subset of items is simply the *sum of the probabilities* of correct answers for those items. Consequently, the score is typically not a whole number.

In addition to providing a mechanism for estimating scores on items that were not administered to every individual, IRT has advantages over raw number-right scoring in the treatment of guessed and omitted items. By using the overall pattern of right and wrong responses to estimate ability, the model does not give credit for correct answers to hard items by low ability students. Omitted items are treated as if the examinee had guessed at random. Raw number-right scoring, in effect, treats omitted items as if they had been answered incorrectly. While this may be a reasonable assumption in a motivated test for older students, this may not always be the case in the ECLS–K, where behavioral or other factors may contribute to a child's inability to complete all items.

4.2 Item Response Theory Estimation Using PARSCALE

The PARSCALE (Muraki & Bock 1991) computer program computes marginal maximum-likelihood estimates of IRT parameters that best fit the responses given by the test takers. The procedure calculates "a", "b", and "c" parameters for each test item, iterating until convergence within a specified level of accuracy is reached. Comparison of the IRT-estimated probability with the actual proportion of correct answers to a test item for examinees grouped by ability provides a means of evaluating the appropriateness of the model for the set of test data for which it is being used. A close match between the IRT-estimated curves and the actual data points means that the theoretical model accurately represents the empirical data.

As indicated earlier, a longitudinal growth study by its very nature consists of subpopulations defined by differing ability levels. That is, after all the kindergarten and first-grade assessments had been completed (four rounds, counting fall and spring administrations) there are four recognizable subpopulations of different ability levels, which are tied to the time of testing. For example, the fall-kindergarten subpopulation will have, on average, a lower expected level of performance than that found in each of the remaining three followups. Similarly the average performance of the fall-first graders will be lower than that of the same children the following spring.

When the first round of kindergarten data was collected in fall 1998, relatively few children were routed to the middle-level second-stage forms and even fewer to the high-level second-stage forms. Thus, there were not enough data on the most difficult items to obtain stable item parameter estimates. As the children were retested in spring-kindergarten and in the fall and spring of first grade the following year, more and more data were collected that could be used to stabilize the estimates for the middle- and then the high-level second-stage items. As each round of data became available, item responses were pooled and parameters re-estimated. The pooling of all time points and re-estimating the item parameters, of course, can lead to a remaking of history in a longitudinal study where intermediate reports are published before all the data from all the time periods are available. That is, fall- and spring-kindergarten scores that have been reported and analyzed might later be modified somewhat when first grade data became available. The use of all data points over time, however, is the preferable method because it is the one method that can provide stable estimates of both the item traces and latent trait scores throughout the entire ability distribution. This procedure was used in the vertical equating that was carried out for National Education Longitudinal Study: (NELS:88) (Rock et al. 1995) and for High School and Beyond (Rock et al., 1985, Rock & Pollack 1987).

A strength of the PARSCALE and other Bayesian approaches to IRT is that they can incorporate information about the ability distribution (i.e., the round of data collection from which an observation is taken) in the ability estimates. This is particularly crucial for measuring change in longitudinal studies. It provides an acceptable way of coping with "perfect" (i.e., all correct scores). For example, a few very advanced children who took the high-level mathematics form in spring-first grade might get all the items correct. These children, while gifted, may not get perfect scores when they eventually are tested on a harder set of items in later grades. Will this mean that they are less knowledgeable in third grade than in first grade? Probably not. Pooling all time points, which amounts to pooling all the items as well as people (in a sense pooling all available information), and recomputing all the item parameters using Bayesian priors reflecting the ability distributions associated with each particular round, provides for an empirically based shrinkage to more reasonable item parameters and ability scores (Muraki & Bock 1991). The fact that the total item pool is used in conjunction with the Bayesian priors leads to shrinking back the extreme item parameters, as well as the perfect scores, to a more reasonable quantity, which in turn allows for the potential of some gains even in the uppermost tail of the distribution. Each of the rounds of data collection in kindergarten and first grade is treated as a separate subpopulation with its own ability distribution. The amount of shrinkage is a function of the distance from the subgroup means and the relative reliability of the score being estimated. Theoretically this approach has much to recommend it. In practice, it has to have reasonable estimates of the difference in ability levels among the subpopulations in order to incorporate realistic priors. Essentially, the scales are determined by the linking items, and the initial prior means for the subgroups are in turn determined by the differential performance of the subpopulations on these linking items. For this reason the item pool has been designed to have an overabundance of items linking forms. This approach, using adaptive testing procedures combined with Bayesian procedures that allow for priors on both ability distributions and on the item parameters, is needed in longitudinal studies to minimize ceiling and floor effects.

A multiple group version of the PARSCALE computer program (Muraki & Bock 1991) that was developed for NAEP allows for both group ability priors and item priors. A publicly available multiple group version of the BILOG (Mislevy & Bock 1982) computer program called BIMAIN (Muraki & Bock 1987, 1991) has many of the same capabilities for dichotomously scored items only. Since the PARSCALE program was applied to dichotomously scored items in the ECLS–K vertical scaling, its estimation procedure is identical to the multiple group version of BILOG or BIMAIN. PARSCALE uses a marginal maximum likelihood estimation approach and, thus, does not estimate the individual ability scores when estimating the item parameters but assumes that the ability **distribution** is known for each subgroup. Thus, the posterior distribution of item parameters is proportional to the product of the likelihood of observing the item response vector, based on the data and conditional of the item parameters and subgroup membership, and the assumed prior ability distribution for that subgroup. More formally, the general model in terms of item estimation is the same as that used in NAEP and described in some detail by Yamamoto and Mazzeo (1992, p. 158) as follows:

$$\begin{aligned} L(\beta) &= \prod_g \prod_{j:g} \int_0 P(x_{j:g} \mid \theta, \beta) f_g(\theta) d(\theta) \\ &\approx \prod_g \prod_j : g \Sigma_k P(x_{j:g} \mid \theta = X_k, \beta) A_g(X_k). \end{aligned}^{1} \qquad (4.2)$$

In equation (4.2), $P(x_{j:g} \mid \theta, \beta)$ is the conditional probability of observing a response vector $x_{j:g}$ of person j from group g, given proficiency θ and vector of item parameters $\beta = (a_1, b_1, c_1, \ldots, a_j, b_j, c_j)$, and $f_g(\theta)$ is a population density for θ in group g. Prior distributions on item parameters can be specified and used to obtain Bayes modal estimates of these parameters (Mislevy 1984). The proficiency densities can be assumed known and held fixed during item parameter estimation or can be estimated concurrently with item parameters.

The $f_g(\theta)$ in (1) are approximated by multinomial distributions over a finite number of quadrature points, where X_k 2 for $k = 1,\ldots,q$, denotes the set of points and $A_g(X_k)$ are the multinomial probabilities at the corresponding points that approximate $f_g(\theta)$ at $\theta = X_k$. If the data are from a single population with an assumed normal distribution, Gauss-Hermite quadrature procedures provide an optimal set of points and weights to best approximate the integral in (1) for a broad class of smooth functions. For more general f or for data from multiple populations with known densities, other sets of points (e.g., equally spaced points) can be substituted, and the values of $A_g(X_k)$ may be chosen to be the normalized density at point X_k (i.e., $A_g(X_k) = f_g(X_k)/\sum_k f_g(X_k)$).

Maximization of $L(\beta)$ is carried out by an application of an EM algorithm (Dempster, Laird & Rubin 1977). When population densities are assumed known and held constant during estimation, the algorithm proceeds as follows. In the E step, provisional estimates of item parameters and the assumed multinomial probabilities are used to estimate expected sample sizes at each quadrature point for each group (denoted $\hat{N}_{gk}$), as well as over all groups (denoted $\hat{N}_k = \sum_g \hat{N}_{gk}$). These same provisional estimates are also used to estimate an expected frequency of correct responses at each quadrature point for each group (denoted $\hat{r}_{gik}$), and over all groups (denoted $\hat{r}_{ik} = \sum_g \hat{r}_{gik}$). In the M step, improved estimates of the item parameters, β, are obtained using maximum likelihood by treating the $\hat{N}_{gk}$ and $\hat{r}_{ik}$ as known, subject to any constraints associated with prior distributions specified for β.

The user of the multiple group version of PARSCALE has the option of fixing the priors on the ability distribution or allowing the posterior estimate to update the previous prior and combine with the data based likelihood to arrive at a new set of posterior estimates after each major EM cycle. If one wishes to update on each cycle, one can continue to constrain the priors to be normal or their shape can be allowed to vary. The ECLS–K approach was to allow for updating the prior but with the normality assumption. The smoothing that came from the updated normal priors led to less jagged looking ability score distributions and did not tend to overfit the item parameters. Lack of fit in the item parameter distribution would simply be absorbed in the shape of the ability distribution if the updated ability distribution were allowed to take any shape. A similar procedure was used in estimating the item parameters in the National Adult Literacy Study (NALS).

It should be remembered that the solution to equation 4.2 finds those item parameters that maximize the likelihood across all four rounds. The present version of the multiple group PARSCALE only saves the subpopulation means and standard deviations and not the individual expected *a posteriori* (EAP) scores. The individual EAP scores, which are the means of the posterior distributions of the latent variate, were obtained from the C-Group conditioning program, which uses the gaussian quadrature procedure. This variation is virtually equivalent to conditioning (e.g., see Mislevy et al. 1992) on a set of "dummy" variables defining which ability subpopulation an observation comes from. The one difference is that the group variances are not restricted to be equal as in the standard conditioning procedure.

Conditional independence is an assumption of all IRT models, but as Mislevy, et al. (1992) point out, not likely to be generally true. However, if one thinks of IRT-based scores as a summarization of essentially the largest latent factor underlying a given item pool, then small violations are of little significance. To insure that there were no substantive violations of this assumption, factor analyses were carried out on the field test forms to confirm that there was a large dominant factor underlying each content area. In addition, all item traces were inspected to insure a good fit throughout the ability range. More importantly estimated proportions correct by item by grade were also estimated in order to insure that the IRT model was both reproducing the actual percent correct (P+) for each item and there was no systematic bias in favor of any particular grade. Since the item parameters were estimated using a model that maximizes the goodness of fit across the rounds, one would not expect much difference here. No systematic bias was found for any particular grade. Appendices D-1 to D-3 list the IRT item parameters for the three subject areas. They also show the actual proportion correct for test takers who answered each item, the proportion correct predicted from the IRT model, and the difference.

5. PSYCHOMETRIC CHARACTERISTICS OF THE ECLS–K DIRECT COGNITIVE BATTERY

This chapter will document the direct cognitive test results for the four rounds of testing in kindergarten and first grade. Note that numbers of observations in some of the tables in this chapter may differ slightly from number of cases in the ECLS–K public release files. These analyses were carried out prior to final determination of cases eligible for the public release files, and a few cases were deleted from the files. There are also small inconsistencies in numbers within tables, most often because a few children answered enough items in the routing section to receive a test score, but no items in a second-stage form.[3]

5.1 Motivation and Timing

An important issue in a low-stakes testing situation is motivation: whether the test results really represent the best efforts of the test takers. There are several pieces of evidence to support the conclusion that the ECLS–K participants were motivated to try their best. Field interviewers reported that children generally enjoyed the testing experience, took it seriously, and were cooperative. At the end of each testing session, assessors assigned a rating of each child's motivation, cooperation, and attention. Tables 5-1 to 5-3 show the distribution of these ratings in each round of testing.

These results show that assessors found the majority of children to be motivated, cooperative, and attentive during the testing sessions. Nearly all children were perceived as cooperative (any of the highest three ratings) at all rounds of testing. Motivation and attentiveness improved slightly between kindergarten and first grade, with over 90 percent of first graders rated in the highest three categories. Another indication of motivation is the very small number of chance-level scores in the tables for the second-stage test forms. This suggests that children were putting effort into their responses rather than responding at random.

There were no time limits on test sections; children were able to proceed at their own speed. Tests were discontinued only if children seemed unable or unwilling to continue. This approach resulted in scorable tests for almost all of the children who started a testing session. As the tables in the sections that follow report, only a very small number of children answered too few items for scores to be calculated.

---

[3] Tests were scored if there were at least 10 items answered in the routing test and second-stage level test combined

For each of the three content domains, the performance of the two-stage procedures, reliabilities, score statistics, and analysis of differential item functioning (DIF) will be presented. First, an expanded explanation and interpretation of DIF is in order.

Table 5-1.—Child's overall motivation level during the assessment

Category	Round 1 N=19,045	Round 2 N=19,884	Round 3* N=5,253	Round 4 N=16,684
Very Low: Child doesn't try or attempt many items, even with encouragement.	1.7%	1.6%	1.0%	1.2%
Low: Child frequently says "I don't know" without even trying, consistent encouragement needed.	9.9%	10.4%	7.5%	8.1%
Average: Child works on most items, says "I don't know" or refuses to answer items after s/he has begun doing some work or after making some attempt to figure the item out.	48.5%	44.5%	44.9%	39.5%
High: Child tries or attempts every item, including some of the most difficult.	29.8%	30.7%	32.5%	31.4%
Very High: Child tries or attempts every item, even the most difficult, appears interested in all the items, may need encouragement to move on to other items.	10.0%	12.9%	14.2%	19.9%
Very Low + Low	11.6%	11.9%	8.5%	9.3%
Average + High + Very High	88.4%	88.1%	91.5%	90.7%

*Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

Table 5-2.—Child's overall cooperation during the assessment

Category	Round 1 N=19,046	Round 2 N=19,884	Round 3* N=5,253	Round 4 N=16,684
Very Uncooperative: Child repeatedly refuses to comply.	1.1%	0.6%	0.4%	0.8%
Uncooperative: Child complies at least 50 percent of the time.	2.7%	2.0%	1.5%	1.3%
Matter of Fact: Child complies at least 75 percent of the time.	22.7%	23.5%	22.1%	23.2%
Cooperative: Child complies with MOST (80-90 percent) requests and directives.	53.2%	49.6%	49.9%	43.5%
Very Cooperative: Child complies with ALL requests and directives in first request.	20.3%	24.3%	26.1%	31.1%
Very Uncooperative + Uncooperative	3.8%	2.6%	1.9%	2.2%
Matter of Fact + Cooperative + Very Cooperative	96.2%	97.4%	98.1%	97.8%

*Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

Table 5-3.—Child's overall attention level during the assessment

Category	Round 1 N=19,046	Round 2 N=19,884	Round 3* N=5,253	Round 4 N=16,684
Unable to Attend: Child needs ongoing redirection to the task.	0.6%	0.6%	0.3%	0.3%
Difficulty Attending: Child is distracted easily and often requires redirection.	13.6%	11.4%	8.0%	9.4%
Attentive: Child attends the majority of the time, when distracted child returns to task with redirection.	43.3%	37.9%	37.9%	35.7%
Very Attentive: Child may momentarily be distracted but is able to return to the task on his/her own.	31.0%	33.9%	35.2%	32.1%
Complete and Full Attention: Child is able to ignore any distractions.	11.5%	16.3%	18.7%	22.5%
Unable to Attend + Difficulty Attending	14.2%	12.0%	8.3%	9.7%
Attentive + Very Attentive + Complete and Full Attention	85.8%	88.0%	91.7%	90.3%

*Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

5.2 Differential Item Functioning

DIF as defined here attempts to identify those items showing an unexpectedly large difference in item performance between a focal group (e.g., Black students) and a reference group (e.g., White students) when the two groups are "blocked" or matched on their total score. It should be noted that any such strictly internal analysis (i.e., without an external criterion) cannot detect bias when that bias pervades all items in the test (Cole & Moss 1989). It can only detect differences in the relationships among items that are anomalous in some group in relation to other items. In addition such approaches can only identify the items where there is unexpected differential *performance*, they cannot directly imply *bias*. A determination of bias implies not only that differential performance on the item is related to subgroup membership but also that the difference is *unfairly* associated with subgroup membership. That is, the difference is due to an attribute *not related* to the construct being measured. As Cole and Moss (1989) point out, items so identified must still be interpreted in light of the intended meaning of the test scores before any conclusion of bias can be drawn. It is not entirely clear how the term item bias applies to academic achievement measures given to students with different patterns of exposure to content areas. For example, some students may be in schools where there is more emphasis on life science topics in kindergarten, while others may begin with units on physical science. Both groups may have similar total scores but for one group the life science items may be differentially difficult while the reverse is true for the other group. It is the Educational Testing Service's (ETS's) practice to carry out DIF analysis on all tests it designs in order to detect test items with differential performance for subgroups defined by gender and ethnicity.

The DIF program was developed at ETS (Holland and Thayer 1986) and was based on the Mantel-Haenszel odds-ratio (Mantel and Haenszel 1959) and its associated chi-square. Basically, the Mantel-Haenszel (M-H) procedure forms odds ratios from two-way frequency tables. In a 20-item test, 21 two-way tables and their associated odds-ratios can be formed for each item. There are potentially 21 of these tables for each item since there will be one table associated with each total score from 0 to 20. The first dimension of each table is groups (e.g., Whites vs. Blacks), and the remaining dimension is passing versus failing on a given item. Thus, the question that the M-H procedure addresses is whether or not members of the reference group (e.g., Whites), who have the same total score as members of the focal group (e.g., Blacks), have the same likelihood of passing the item in question. While the M-H statistic looks at passing rates for two groups while controlling for total score, no assumption need be made about the shape of the total score distribution for either group. The chi-square statistic associated with the M-H procedure tests whether the average odds-ratio for a test item, aggregated across all 21 score levels differs from unity (i.e., equal likelihood of passing).

The M-H procedure provides a statistical test of whether or not the average odds-ratio significantly departs from unity for each item. If the probability is .05 or lower, then one could say that there is statistical evidence for DIF on the item in question. The problem with this interpretation is two-fold. First, a very large number of statistical tests are being performed, one for each item for each pair of subgroups, so low probabilities will be found occasionally even if no DIF is present. Second, if there are two relatively large samples involved, statistical significance will be guaranteed.

Given these reservations, ETS has developed an "effect size" estimate that is not sample size dependent. Associated with the effect sizes is a letter code that ranges from "A" to "C." It is ETS's experience that effect sizes of 1.5 and higher have practical significance. Effect sizes of this magnitude, and which are statistically significant, are labeled with a "C." Items labeled "A" or "B" either do not show statistically significant differential functioning for the two groups being compared or have differences that are too small to be important. Test development experts inspect items that are characterized by such large DIF properties and in some cases are able to identify the reason, other than bias, for the DIF.

The negative numbers in some cells of the DIF tables for mathematics and general knowledge in sections 5.4.4 and 5.5.4 indicate that more C-DIF items favor the focal group (females or minority groups) than the reference group (males or White children) for these cells.

5.3 Reading Test

5.3.1 Samples and Operating Characteristics

Table 5-4A presents sample counts and operating characteristics of the adaptive test forms in reading. The small sample size reported at round 3 in table 5-4A reflects the fact that only a subsample of the fall-first-grade longitudinal cohort was assessed at this point in time. The line labeled "Too few items" refers to the number of children who did not attempt a sufficient number of reading items to generate a reliable score. Scores were calculated only for children who attempted at least ten items. Only a fraction of one percent of the kindergartners and almost none of the first-graders were unable or unwilling to complete enough test items to receive a score.

Table 5-4A.—Reading test: samples and operating characteristics*

Characteristics	Round 1	Round 2	Round 3	Round 4
Sample N	17,630	18,944	5,054	16,340
Too few items	44	19	0	2
Number taking low form	13,355 (76%)	6,521 (34%)	1,062 (21%)	618 (4%)
Number taking middle form	3,620 (21%)	8,906 (47%)	2,334 (46%)	2,371 (15%)
Number taking high form	654 (4%)	3,517 (19%)	1657 (33%)	13,351 (82%)
% perfect score routing test	.3%	1.7%	4.9%	23.6%
% perfect score low form	0%	.1%	.4%	1.6%
% perfect score middle form	0%	0%	0%	0%
% perfect score high form	0%	.2%	0%	0%
% less than chance routing test	22.6%	3.7%	2.1%	.3%
% less than chance low form	.9%	.5%	.2%	6%
% less than chance middle form	.5%	.3%	.1%	.1%
% less than chance high form	.5%	1.7%	2.3%	.4%

* The table excludes language minority children who did not achieve the English OLDS cutting score. Round 3 is a subset of approximately 30 percent of the full ECLS-K sample.

The percentages taking the various second-stage forms in reading followed the expected distributions based on the cut points determined by simulations of the field test data. That is, in round 1 about three-quarters of the children were assigned the low second-stage form based on their routing test performance. In rounds 2 and 3, the largest percentages were assigned the middle-level form. By spring-first grade, round 4, more than three-quarters of the students took the highest level of the second-stage forms.

More important than the routing percentages matching the intended targets is whether the cutting scores succeeded in routing children to a second stage test of an appropriate level of difficulty. The percentages of perfect and less than chance scores in table 5-4A demonstrate that the two-stage test design accomplished its objective of avoiding floor and ceiling effects. The percentages of perfect scores were all close to zero with exception of the round 4 routing test. Although about 23 percent of the children had perfect scores on the routing test in round 4, the main function of the routing test is to make a proper assignment to the correct second-stage form. The children were then scored on the combination of their first- and second-stage items combined. Since there was no ceiling effect problem in the high-level second-stage form (no perfect scores at all when the supplementary items were included), the perfect routing test scores do not have the potential to create a ceiling effect. Table 5-4A also shows little or no evidence of a floor effect when both first- and second-stages are combined to compute ability levels and scale scores. While 22.6 percent scored below chance on the routing test in round 1, these children were routed to the low-level second-stage form where more than 99 percent of them were able to respond at or above the chance level. Again, their scores reflected performance on the combined set of routing and second-stage items.

5.3.2 Reliabilities

Table 5-4B presents the internal consistency (alpha) coefficients for the routing test and the second-stage forms. These classical estimates of reliability of the routing test are quite high for a 20-item test, in the middle to high 80s for each round. The internal consistency coefficients for the second-stage forms were generally lower due to the restriction in range among the children sent to the various second-stage forms. Since the children taking each of these forms are a more homogeneous group with respect to reading performance, the score variance, and thus the alpha coefficient, are lower than they would have been if the whole sample of children had taken each set of items. Only for the high-level second-stage form, which had much greater variance than did the other forms, did the alpha coefficients approach or exceed .90.

Table 5-4B.—Reading test: reliabilities and mean score gains

Characteristics	Round 1	Round 2	Round 3*	Round 4
Alpha routing	.86	.88	.88	.86
Alpha low form	.69	.69	.71	.72
Alpha middle form	.70	.72	.74	.78
Alpha high form	.90	.88	.93	.92
Reliability of theta	.93	.95	.96	.97
Proficiency level 1 reliability	.83	.79	.77	.78
Proficiency level 2 reliability	.76	.76	.73	.70
Proficiency level 3 reliability	.72	.76	.76	.78
Proficiency level 4 reliability	.78	.77	.80	.78
Proficiency level 5 reliability	.60	.69	.73	.73
Mean (SD) routing test (maximum = 20)	5.83 (3.98)	10.07 (4.17)	11.85 (4.28)	16.48 (3.55)
Mean (SD) theta	.00 (.81)	.87 (.76)	1.23 (.75)	2.07 (.67)
Mean (SD) scale score (maximum = 92)	22.67(8.58)	32.47 (10.85)	37.97 (12.67)	54.77 (14.17)

*Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

The most appropriate estimate of reliability for the full reading test is based on the Item Response Theory (IRT) theta scores. Inspection of the table 5-4B indicates that the reliability of the theta scores (ability estimates) ranges from .93 to .97. These are more appropriate estimates since they reflect the internal consistency for performance on the combined first- and second-stage sections and for the full range of variance found in the sample as a whole. One would expect the reliability of the scale scores to be similar to that of the thetas since they are a nonlinear transformation of the theta scores.

Split-half reliabilities are shown for the clusters of items that define each of the proficiency levels in the reading test. These are generally in the high 70s, which is quite high given that each cluster contained only four items. One would expect them to be internally consistent, however, since they were selected to be criterion-referenced marker items that are measuring essentially the same skill at the same difficulty level. The lower reliabilities for proficiency level 5, especially in the kindergarten rounds, reflect the fact that the routing test that contained these items was discontinued prior to this cluster for children who were not able to succeed at the easier tasks. Thus, the restricted variance for those who did answer the items resulted in a lower estimate of reliability at level 5 than for the clusters answered by all test takers.

The proficiency-level reliabilities in the table apply to the use of the dichotomous (0 or 1) observed mastery scores. These scores are not the generally recommended approach to defining proficiency or mastery levels since not everyone answers all the clusters of items. For the continuous proficiency-level probability scores, which are recommended for analysis, the reliability of the theta is an appropriate measure of internal consistency.

5.3.3 Score Gains

Inspection of the reading means by rounds suggests that there is both rapid and differential growth between adjacent rounds. That is, the maximum gains in reading performance occur in first grade between rounds 3 and 4 of data collection. Gains nearly as large in terms of standard deviation units occur during the kindergarten year, round 1 to round 2, with somewhat smaller gains found over the summer period, round 2 to round 3.

5.3.4 Differential Item Functioning

As described earlier, DIF refers to a statistical procedure for identifying the tendency for some population subgroups to do comparatively worse on some items compared with a reference subgroup, even though they have similar total scores. Each focal group, for example, racial/ethnic minority groups, is compared with a reference group (e.g., White children). The fact that an item is identified by the DIF procedure does not mean that the item is necessarily unfair to any particular group. The DIF procedure is merely a statistical screening step that indicates that the item is behaving somewhat differently for one or more subgroups. In an achievement test this could simply result from differences in curriculum or other reasons for differential exposure to some particular knowledge. Thus, the formal DIF analysis is the first step in a two-step screening procedure. As indicated in the discussion of DIF in chapter 3, C-DIF items show sufficient gaps in performance to alert the test constructor to further investigate the item content for evidence that the item may be measuring some extraneous dimension not consistent with the test framework. Items that attain C-level DIF in favor of the majority group are routinely submitted for a review of the content by a standing committee in which the relevant minority group is represented. This is the second stage in the screening procedure. If the committee decides that the item content is measuring important content that is consistent with the test framework and does not contain language or context that would be unfair to a particular group, the item is kept in the test. If the committee finds otherwise, the item is either modified or removed from the test.

Table 5-4C summarizes the results of the DIF analysis. Each row presents the net number of C-DIF items that favor the reference group. For example, the results of White versus Black DIF analysis in round 1 showed six items favoring the focal group (Black children) and nine favoring the reference group (White children) for a net count of three in favor of the reference group.

Table 5-4C.—Reading test: differential item functioning

Characteristics	Round 1	Round 2	Round 3	Round 4
Net # C-DIF items favoring Male vs. Female	1	0	1	1
Net # C DIF items favoring White vs. Black	3	3	3	6
Net # C-DIF items favoring White vs. Hispanic	3	5	5	4
Net # C-DIF items favoring White vs. Asian	3	6	2	5
Net # C-DIF items favoring high SES vs. low SES	2	3	3	4

Inspection of the DIF results in table 5-4C shows net DIF counts ranging from zero to six for the various groups being compared. In all cases where the DIF occurred against the minority group, the items were submitted to the fairness committee for review and all the items were passed by the committee. In other words, a judgment was made that the difference in performance was due to skill differences consistent with the test specifications and not due to factors that would unfairly bias the item against the subgroup. It should be kept in mind that there are 72 reading items in the test forms for kindergarten and 92 for first grade. With five sets of comparison groups and four rounds of data, more than 1,500 comparisons are made in an attempt to statistically identify items showing DIF, so chance alone could account for some of the findings. In an achievement test covering reading development where children are growing very fast but at quite different rates, one might identify DIF at one time point and then see it reduced or go away later in the children's development. However, more DIF items were identified in the reading test than in mathematics or general knowledge. This suggests that certain subgroups may not all follow the same pattern and/or rate of development with respect to their reading performance and language arts acquisition.

5.4 Mathematics Test

5.4.1 Samples and Operating Characteristics

Table 5-5A presents sample information and operating characteristics for the mathematics test forms for the kindergarten and first grade administrations. Inspection of table 5-5A shows allocations to second-stage forms that are quite similar to those of the reading test in fall-kindergarten and spring-first grade. That is, in round 1 about three-quarters of children were routed to the low-level second-stage form, while by round 4 more than three-quarters advanced to the high-level form. In rounds 2 and 3, fewer children received the middle-level form than in reading, with more at the extremes. This may be due to differences in curriculum emphasis in different schools. While most kindergarten and first-grade classes probably put their major emphasis on the development of reading skills, there may be more variation in the amount of attention given to mathematics concepts and skills prior to first grade. Again, the important point here is not matching the anticipated routing percentages but matching the test form to each child's ability level. The percentages of perfect and less than chance scores in the table show that the routing algorithms were successful in avoiding floor effects (very few less than chance scores in second-stage low form) and ceiling effects (even fewer perfect scores in second-stage high form).

Table 5-5A.—Mathematics test: samples and operating characteristics*

Characteristics	Round 1	Round 2	Round 3	Round 4
Sample N	18,641	19,657	5,226	16,647
Too few items	21	15	0	2
Number taking English version	17,615	18,925	5,049	16,336
Number taking Spanish version	1,021	724	177	305
Number taking low form	14,380 (77%)	8,444 (43%)	1,353 (26%)	1,097 (7%)
Number taking middle form	3,123 (17%)	6,169 (31%)	1,521 (29%)	2,317 (14%)
Number taking high form	1,136 (6%)	5,042 (26%)	2,351 (45%)	13,233 (79%)
Percent perfect score routing test	.1%	.4%	1.5%	7.9%
Percent perfect score low form	.1%	.4%	1.0%	2.5%
Percent perfect score middle form	0%	0%	0%	.3%
Percent perfect score high form	0%	0%	0%	.1%
Percent less than chance routing test	15.3%	3.1%	1.6%	.3%
Percent less than chance low form	.9%	.3%	.1%	.3%
Percent less than chance middle form	.1%	0%	0%	0%
Percent less than chance high form	.1%	0%	0%	0%

* The table excludes language minority children who did not achieve the English OLDS cutting score. Round 3 is a subset of approximately 30 percent of the full ECLS-K sample.

Chapter 3 describes the development of a Spanish translation of the mathematics test for children who could not be tested in English but had sufficient fluency in Spanish. Table 5-5A shows the steady decline in the proportion of children who were tested in Spanish. More than two-thirds of the children who received the Spanish mathematics test in fall-kindergarten were able to take the English version of the test by spring of first grade.

As in the reading test, there did not seem to be any significant floor or ceiling effects in the mathematics test. Less than one percent of children received either chance scores or perfect scores when the routing and second-stage forms were combined.

5.4.2 Reliabilities

The internal consistency (alpha) coefficients for the individual mathematics test forms shown in table 5-5B were slightly lower than for reading, reflecting more diversity in curriculum topics that might be expected for mathematics. As with reading, the alpha coefficients for the low and middle second-stage forms were lower than for the routing test, an artifact of the restricted variance found in the second-stage forms. The greater number of test items in the high second-stage form, although it too was given to a selected sample, resulted in alpha coefficients comparable to the routing test (see table 5-5B).

Also similar to the reading test, the reliabilities of the mathematics theta scores were in the mid-90s. These findings suggest quite high reliability given the number of items administered to each child.

The split-half reliabilities for the mathematics proficiency-level clusters were substantially lower than those in the reading test for two reasons. First, the mathematics clusters generally were not as homogeneous with respect to similarity of content and skill demand as was the case with reading. Second, not all children received the complete set of mathematics proficiency items. In the reading test, these clusters were located entirely in the routing test, so children of all skill levels attempted them (unless the routing test was discontinued before the end). This was not the case in mathematics, where some of the proficiency cluster items were located in the second-stage forms. One would expect that the reliabilities of the mathematics cluster scores would be reduced to the extent that the children answering the items were more homogeneous with respect to mathematics achievement. The greater heterogeneity of content of

Table 5-5B.—Mathematics test: reliabilities and mean score gains

Characteristics	Round 1	Round 2	Round 3*	Round 4
Alpha routing	.78	.81	.83	.80
Alpha low form	.70	.66	.66	.71
Alpha middle form	.66	.67	.66	.66
Alpha high form	.80	.80	.83	.82
Reliability of theta	.92	.94	.94	.94
Proficiency level 1 reliability	.41	.27	.26	.26
Proficiency level 2 reliability	.58	.49	.51	.32
Proficiency level 3 reliability	.63	.66	.67	.59
Proficiency level 4 reliability	.54	.63	.66	.63
Proficiency level 5 reliability	.46	.53	.61	.65
Mean (SD) routing test (maximum = 16)	4.54 (2.95)	7.32 (3.27)	8.91 (3.36)	11.78 (2.96)
Mean (SD) theta	-.18 (1.00)	.80 (.95)	1.32 (.94)	2.26 (.84)
Mean (SD) scale score (maximum = 64)	19.30 (7.11)	27.16 (8.74)	32.37 (9.61)	42.78 (9.50)

* Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

the item clusters compared with reading, and the greater homogeneity of the children answering the items in each cluster, would both serve to depress the reliability coefficients.

The recommendation made earlier in the reading discussion to use the children's continuous probabilities of proficiency rather than the dichotomous 0 or 1 proficiency level scores is even more important here than in reading.

5.4.3 Score Gains

The high reliabilities of theta for the overall mathematics test indicates that the test scores would be sensitive measures of growth in mathematics achievement. Growth was fast during the kindergarten and first-grade school years (rounds 1 to 2, and 3 to 4), averaging about a standard deviation in both the scale score and theta metrics during those periods. During the summer between kindergarten and first grade (round 2 to round 3), gains were closer to one-half a standard deviation. These numbers are comparable to the reading results.

5.4.4 Differential Item Functioning

Inspection of the DIF rows in table 5-5C suggests that there is little DIF with the possible exception of the White versus Hispanic contrast, especially in round 4. The negative numbers in the table indicate that for these sets of contrasts the minority groups have more items showing DIF in favor of them than do the reference groups.

Table 5-5C.—Mathematics test: differential item functioning

Characteristics	Round 1	Round 2	Round 3	Round 4
Net # C-DIF items favoring Male vs. Female	-1	0	0	0
Net # C DIF items favoring White vs. Black	0	-1	1	0
Net # C-DIF items favoring White vs. Hispanic	1	4	2	6
Net # C-DIF items favoring White vs. Asian	-1	0	2	4
Net # C-DIF items favoring high SES vs. low SES	1	-1	1	1
Net # C-DIF items favoring English vs. Spanish	3	2	0	4

BEST COPY AVAILABLE

The table of DIF statistics for mathematics has one more set of contrasts than those for the other subject areas: performance for children who took the test in English compared with those who took the Spanish mathematics translation. Examination of the English versus the Spanish DIF statistics for the proficiency level item clusters in round 4 suggests that those who were still taking the mathematics test in Spanish in round 4 are not doing as well as expected on items defining mathematics proficiency levels 2 and 3. Typically, high proportions of children should show mastery at these levels by spring-first grade.

5.4.5 Comparability of Spanish Mathematics Test

Several analyses were undertaken to establish the comparability of scores derived from the English and Spanish versions of the mathematics test. Analysis of DIF (table 5-5C), a comparison of actual and predicted item performance, and IRT fit statistics support the conclusion that the two versions of the test are functioning in a similar manner. In addition, analysis of score gains for children who made the transition from the Spanish to the English test between spring-kindergarten and spring-first grade finds gains in mathematics that are comparable in size.

Table D-4 in appendix D presents the results of an analysis of actual versus predicted item performance for the English and Spanish versions. Data from all four rounds of data collection were combined for stability of estimates, since the number of Spanish test takers was small in rounds 3 and 4. The columns of "actual P+" show the percent correct for children answering each item in each version of the test. The "predicted P+" show the mean probabilities of correct answers based on IRT estimates of item parameters and children's ability estimates. Deviations of actual from predicted performance were very small for all items in the English version of the test. This is at least partly due to the disproportionate number of English test takers, about 97 percent, in the IRT calibration. For the Spanish version, the deviations were substantially larger. For 12 of the 64 test items, the discrepancy between actual and predicted performance was 10 percentage points or more. According to an expert in international literacy testing (K. Yamamoto, personal communication, September-November, 2001), discrepancies of up to 10 percentage points on tests translated into different languages are considered inconsequential. Of the 12 test items that exceeded this limit, half were for items in the high second-stage form of the mathematics test. Fewer than 200 children took the high second-stage level form in Spanish, which may have resulted in somewhat unstable estimates. For five of these six high-form items, the children taking the Spanish test actually did *better* than predicted on the basis of the IRT-based estimates. The six routing, low, and middle form items that had discrepancies exceeding 10 percentage points were evenly divided: on three of them, actual performance was better than expected, and on the other three, worse than expected. Averaging the discrepancies over all items, without regard to sample size, resulted in no difference for low second-stage form items (more than 80 percent of the Spanish tests routed to the low second-stage form). On the routing test, the average underperformance was 3 percentage points, while the middle and high forms both showed performance exceeding expectations by an average of 2 percentage points.

A more direct look at statistics from a strictly IRT-based perspective produced similar results. Graphs of item response functions were examined for fit of English and Spanish test groups. In addition, Bayesian estimates of fit statistics are shown in table D-5 in appendix D. Mean deviations correspond to the differences between observed and projected proportion correct while taking account of the different ability distributions of the two groups. If an item is more difficult for one group of students, the value of the mean deviation will be negative. This statistic shows both the direction and magnitude of the deviation. The magnitude of deviations should be interpreted along with the balance of positive and negative deviations for each subgroup.

Within a single item, positive and negative mean deviations for a subgroup may cancel each other out when summed over the range of the ability distribution. The root mean squared deviations take into account the absolute amount of deviation regardless of direction. The absolute values of mean deviation and root mean squared deviation for items in the table are quite similar. This similarity suggests that deviations are largely due to a uniform shift of item characteristics. Although a few items showed mean deviations greater than .10, the large number of items administered to each student, and the balance of positive and negative mean deviations, means that these deviations would have very little impact on overall ability estimates.

The three methods of assessing comparability, DIF analysis, actual versus predicted performance, and analysis of fit statistics, all produced very similar results. Approximately the same items were identified as having small but discernible differences in performance. The few items that tended to be differentially more difficult on the Spanish mathematics test tended to have more verbiage, while the items that were differentially easier tended to rely more heavily on numbers. But the great majority of test items performed similarly in the two versions. While it is possible that children's scores on the Spanish mathematics test may differ slightly from what they might have been had they been able to be tested in English, the evidence suggests that such discrepancies would be small.

Additional analyses were undertaken to determine whether the language of the test might affect measurement of gain. There has been some concern that Spanish-speaking children who fail the English OLDS and take the mathematics test in Spanish but then in the succeeding round pass the screener and are tested in English may show a lack of gain in knowledge. Table 5-5D speaks to this concern.

Table 5-5D.—Performance of children who took Spanish mathematics test in round 2 and English mathematics test in round 4

Characteristics	Mean	N	Standard deviation	Standard error mean
Mathematics theta round 2	-.056	319	.884	.050
Mathematics theta round 4	1.697	319	.827	.046

Inspection of table 5-5D suggests just the opposite. That is, those children who took the mathematics test in Spanish in round 2 (spring-kindergarten) and then in English in round 4 (spring-first-grade) gained almost 2 standard deviations (in theta units). This is equal to or greater than the average gain for the general population. The children who continued to take the mathematics test in Spanish through round 4 had mean scores and gains from round 2 to round 4 that were very similar to the statistics for the children who moved from the Spanish to the English version during this period of time. This supports the idea that the Spanish translation of the mathematics test is functioning in a manner similar to the English version.

5.5 General Knowledge Test

5.5.1 Samples and Operating Characteristics

Table 5-6A presents sample information and operating characteristics for the general knowledge test at each of four rounds.

Table 5-6A.—General knowledge test: samples and operating characteristics*

Characteristics	Round 1	Round 2	Round 3	Round 4
Sample N	17,571	18,910	5,044	16,328
Too few items	23	25	0	9
Number taking low form	12,286 (70%)	9,323 (49%)	1,794 (36%)	3,437 (21%)
Number taking high form	5,285 (30%)	9,587 (51%)	3,250 (64%)	12,891 (79%)
Percent perfect score routing test	.6%	2.1%	4.4%	9.5%
Percent perfect score low form	0%	0%	0%	0%
Percent perfect score high form	0%	0%	.1%	.2%
Percent less than chance routing test	8.1%	6.8%	3.4%	1.7%
Percent less than chance low form	1.0%	.9%	.8%	.4%
Percent less than chance high form	0%	0%	0%	0%

* The table excludes language minority children who did not achieve the English OLDS cutting score. Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

As in the case of reading and mathematics, participation rates in the general knowledge domain test were high. Since growth in this content area was expected to be less than in reading and mathematics, there were only two, instead of three, second-stage forms. As planned, almost three-quarters of fall-kindergartners were routed to the low level second-stage form, and slightly more than three-quarters of spring-first graders to the high-level form. There appear to be no floor or ceiling effects, with less than one percent of test takers receiving chance or perfect scores on the full set of items received. This confirms that the cut points were successful in routing each child to a second stage test form of appropriate difficulty.

5.5.2 Reliabilities

Inspection of the data in table 5-6B shows alpha coefficients for the routing test that are comparable to those for reading and mathematics, with somewhat lower alphas for the second-stage forms due to the restricted variance of the test takers within each. The reliability of theta is slightly lower than for the other tests but still very close to the desired target reliability of .90. At any rate it has sufficient reliability to reasonably measure change at both the individual and group level.

Table 5-6B.—General knowledge test: reliabilities and mean score gains

Characteristics	Round 1	Round 2	Round 3	Round 4
Alpha routing	.79	.79	.79	.78
Alpha low form	.72	.70	.68	.69
Alpha high form	.64	.68	.71	.74
Reliability of theta	.88	.89	.89	.89
Mean (SD) routing test (maximum = 12)	4.76 (2.96)	6.32 (3.01)	7.31 (2.92)	8.49 (2.66)
Mean (SD) theta	-.33 (.56)	.01 (.56)	.24 (.56)	.53 (.57)
Mean (SD) scale score (maximum = 51)	22.10 (7.44)	26.81 (7.89)	30.02 (7.92)	34.00 (7.74)

Because of the heterogeneity of content of the test and diversity of curriculum in the areas of science and social studies, no hierarchical proficiency levels were defined for general knowledge.

5.5.3 Score Gains

It is interesting to note that gains from a full year of schooling (fall to spring, in both kindergarten and first grade) in terms of standard deviation units on general knowledge appear to be considerably less than those that were demonstrated in both reading and mathematics. Also, there is less differential in growth rates exhibited between adjacent rounds than in reading and mathematics. The rate of growth during the summer between kindergarten and first grade is closer to the growth during the school year intervals than was found in reading and mathematics. It would appear that the general knowledge test is measuring information that is not necessarily included in most kindergarten and first-grade curricula but is associated more with the child's out-of-school experiences.

5.5.4 Differential Item Functioning

DIF in favor of the reference group was not found on the general knowledge test. The few items identified as having C-DIF were more likely to favor the minority group than the White children.

Table 5-6C.—General knowledge test: differential item functioning

Characteristics	Round 1	Round 2	Round 3	Round 4
Net # C-DIF items favoring Male vs. Female	0	0	0	0
Net # C DIF items favoring White vs. Black	-1	-3	-2	0
Net # C-DIF items favoring White vs. Hispanic	0	0	-2	0
Net # C-DIF items favoring White vs. Asian	2	-2	-1	2
Net # C-DIF items favoring high SES vs. low SES	0	0	0	0

5.6 Intercorrelations of the Direct Cognitive Measures Within Rounds 1 to 4

Evidence for the construct validity of the direct measures of children's achievement can be generated by observing certain consistent correlational patterns within and across the rounds. Table 5-7 presents the intercorrelations of the direct cognitive measures by round.

Table 5-7.—Intercorrelations of the direct cognitive measures within rounds 1 to 4

Tests	Reading	Mathematics	General knowledge
Round 1			
Reading	1.00		
Mathematics	.77	1.00	
General Knowledge	.57	.64	1.00
Round 2			
Reading	1.00		
Mathematics	.76	1.00	
General Knowledge	.57	.66	1.00
Round 3*			
Reading	1.00		
Mathematics	.77	1.00	
General Knowledge	.57	.66	1.00
Round 4			
Reading	1.00		
Mathematics	.74	1.00	
General Knowledge	.59	.67	1.00

* 30 percent subsample

Inspection of the intercorrelations among the ability estimates (thetas) indicates that the relationship between the more school-related measures, reading and mathematics, remains relatively stable through the early schooling years and moderately high (.74 to .77). With the exception of round 3, which is a small subsample of the longitudinal cohort, there may be a slight trend toward more specificity of the reading and mathematics skills as evidenced by the slight decreases in their intercorrelations over time. The less school-related measure, general knowledge, also maintains a stable but differential relationship with reading and mathematics. In all four rounds, general knowledge has a consistently higher relationship with mathematics (.64 to .67) than it does with reading (.57 to .59). At these early developmental stages it would seem that reading is somewhat more of a specific skill than is mathematics.

5.7 Test Results by Round and Selected Demographics

Table 5-8 presents the means and standard deviations for the reading IRT theta scores for selected subpopulations by round. Tables 5-9 and 5-10 present the same information on mathematics and general knowledge IRT theta scores respectively. Tables 5-11 through 5-13 provide scale score statistics for reading, math, and general knowledge. Tables 5-14 through 5-23 show mastery rates for the five proficiency levels in reading and five in mathematics.

5.8 Test Item Usage and Item Performance

Appendices D-1 through D-3 present additional information on the reading, mathematics, and general knowledge test items. For each item, the tables show the test form or forms on which it was used, its IRT parameters, and the number of children who responded to the item in each round of testing. The item fit information compares actual item performance for children who took each item with the estimate of the proportion passing based on the IRT model. The difference between actual and predicted percent correct is shown for each item. For the majority of the items the residual differences lie within plus or minus .02, indicating a very close fit between the observed and estimated proportions. Differences larger than this tended to be for items with very low numbers of observations; for example, the few fall-kindergarten children who were routed to the highest reading form or the small number of spring-first-graders who had not yet progressed beyond the low second-stage form.

Table 5-8.—Reading Item Response Theory theta score (range of possible values: -5 to 5)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	0.00	0.81	18,937	0.87	0.76	5,053	1.23	0.75	16,336	2.07	0.67
Male	8,984	-0.07	0.82	9,688	0.80	0.78	2,556	1.15	0.77	8,349	2.00	0.70
Female	8,640	0.07	0.80	9,247	0.95	0.73	2,497	1.31	0.72	7.987	2.14	0.62
White, non-Hispanic	10,433	0.13	0.78	11,073	0.99	0.72	2,935	1.36	0.70	9,435	2.19	0.62
Black, non-Hispanic	2,854	-0.22	0.76	2,968	0.63	0.76	782	1.03	0.72	2,371	1.84	0.71
Hispanic, race specified	1,182	-0.22	0.82	1,315	0.76	0.77	322	1.16	0.70	1,233	1.96	0.66
Hispanic, race not specified	1,195	-0.37	0.77	1,423	0.58	0.78	377	0.86	0.78	1,335	1.82	0.66
Asian	897	0.34	0.90	1,089	1.18	0.76	257	1.52	0.84	1,042	2.25	0.66
Hawaiian or other Pacific Islander	186	-0.11	0.85	202	0.70	0.76	93	0.90	0.69	188	2.00	0.57
American Indian or Alaska Native	354	-0.59	0.74	344	0.44	0.76	126	0.49	0.77	298	1.59	0.69
More than one race, non-Hispanic	476	-0.03	0.85	473	0.84	0.77	152	1.23	0.73	397	2.10	0.67
SES: first quintile	2,594	-0.52	0.67	2,917	0.43	0.72	753	0.75	0.73	2,363	1.67	0.70
SES: second quintile	3,271	-0.21	0.72	3,503	0.70	0.74	925	1.04	0.71	2,796	1.96	0.65
SES: third quintile	3,470	-0.04	0.73	3,686	0.87	0.69	997	1.28	0.67	3,003	2.10	0.58
SES: fourth quintile	3,650	0.18	0.75	3,909	1.05	0.67	1,019	1.42	0.63	3,173	2.23	0.55
SES: fifth quintile	3,880	0.50	0.80	4,152	1.29	0.69	1,159	1.63	0.66	3,642	2.42	0.54
Public school	13,737	-0.08	0.79	14,579	0.81	0.75	3,809	1.18	0.74	12,998	2.03	0.67
Private school	3,888	0.39	0.79	4,358	1.19	0.72	1,042	1.58	0.63	3,279	2.35	0.56

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

BEST COPY AVAILABLE

Table 5-9.—Mathematics Item Response Theory theta score (range of possible values: -5 to 5)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	-0.18	1.00	19,649	0.80	0.95	5,226	1.32	0.94	16,641	2.26	0.84
Male	9,479	-0.19	1.04	10,041	0.80	0.98	2,644	1.31	1.00	8,506	2.27	0.88
Female	9,156	-0.17	0.95	9,606	0.81	0.91	2,582	1.33	0.88	8,135	2.24	0.79
White, non-Hispanic	10,433	0.09	0.94	11,071	1.06	0.87	2,935	1.57	0.86	9,436	2.46	0.78
Black, non-Hispanic	2,855	-0.53	0.88	2,962	0.42	0.89	781	0.97	0.94	2,371	1.86	0.86
Hispanic, race specified	1,588	-0.54	0.97	1,624	0.51	0.94	389	1.14	0.92	1,354	2.08	0.85
Hispanic, race not specified	1,800	-0.80	0.94	1,834	0.29	0.94	486	0.80	0.94	1,518	1.93	0.78
Asian	898	0.24	1.00	1,089	1.11	0.91	256	1.59	0.94	1,042	2.37	0.83
Hawaiian or other Pacific Islander	187	-0.36	0.93	202	0.54	0.86	93	0.96	0.75	188	1.95	0.71
American Indian or Alaska Native	354	-0.76	0.97	345	0.38	0.90	126	0.57	1.02	298	1.85	0.80
More than one race, non-Hispanic	473	-0.17	0.95	472	0.79	0.88	151	1.25	0.90	397	2.27	0.84
SES: first quintile	3,269	-0.86	0.89	3,426	0.20	0.89	895	0.70	0.94	2,572	1.79	0.86
SES: second quintile	3,429	-0.42	0.91	3,607	0.62	0.90	942	1.12	0.90	2,839	2.09	0.84
SES: third quintile	3,546	-0.14	0.87	3,721	0.85	0.83	1,001	1.40	0.81	3,017	2.30	0.75
SES: fourth quintile	3,676	0.11	0.88	3,921	1.06	0.82	1,023	1.57	0.77	3,178	2.47	0.70
SES: fifth quintile	3,893	0.48	0.91	4,161	1.37	0.84	1,158	1.90	0.82	3,644	2.73	0.66
Public school	14,702	-0.27	0.99	15,260	0.73	0.94	3,971	1.27	0.95	13,292	2.21	0.85
Private school	3,934	0.31	0.93	4,389	1.21	0.87	1,043	1.78	0.73	3,286	2.58	0.68
English version of test	17,615	-0.12	0.98	18,925	0.84	0.93	5,049	1.36	0.93	16,336	2.27	0.83
Spanish version of test	1,021	-1.18	0.83	724	-.19	0.89	177	0.51	0.85	305	1.60	0.99

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-10.—General knowledge Item Response Theory theta score (range of possible values: -5 to 5)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,566	-0.33	0.56	18,903	0.01	0.56	5,044	0.24	0.56	16,324	0.53	0.57
Male	8,944	-0.32	0.57	9,665	0.03	0.57	2,550	0.26	0.58	8,341	0.56	0.57
Female	8,621	-0.34	0.55	9,236	-0.01	0.55	2,494	0.22	0.55	7,983	0.51	0.56
White, non-Hispanic	10,407	-0.15	0.51	11,065	0.20	0.50	2,932	0.43	0.50	9,432	0.73	0.49
Black, non-Hispanic	2,840	-0.70	0.50	2,954	-0.35	0.51	780	-0.07	0.48	2,365	0.20	0.52
Hispanic, race specified	1,175	-0.49	0.54	1,312	-0.13	0.53	321	0.10	0.56	1,233	0.29	0.58
Hispanic, race not specified	1,192	-0.63	0.51	1,421	-0.32	0.51	376	-0.15	0.52	1,335	0.17	0.51
Asian	895	-0.50	0.56	1,083	-0.19	0.55	256	0.05	0.59	1,041	0.36	0.62
Hawaiian or other Pacific Islander	186	-0.63	0.53	202	-0.35	0.55	93	-0.19	0.56	188	0.21	0.53
American Indian or Alaska Native	352	-0.68	0.54	344	-0.25	0.50	126	-0.25	0.53	298	0.27	0.55
More than one race, non-Hispanic	471	-0.34	0.51	472	0.03	0.50	151	0.26	0.54	395	0.60	0.47
SES: first quintile	2,583	-0.74	0.48	2,910	-0.39	0.50	753	-0.19	0.49	2,361	0.12	0.51
SES: second quintile	3,258	-0.47	0.51	3,495	-0.11	0.51	923	0.11	0.50	2,793	0.42	0.52
SES: third quintile	3,463	-0.33	0.50	3,680	0.02	0.49	994	0.23	0.52	3,002	0.56	0.49
SES: fourth quintile	3,638	-0.20	0.50	3,901	0.17	0.49	1,018	0.40	0.47	3,169	0.71	0.47
SES: fifth quintile	3,869	0.03	0.52	4,149	0.39	0.51	1,157	0.64	0.47	3,641	0.91	0.48
Public school	13,682	-0.37	0.56	14,545	-0.03	0.56	3,803	0.21	0.56	12,989	0.49	0.57
Private school	3,884	-0.10	0.54	4,358	0.24	0.53	1,040	0.49	0.50	3,277	0.79	0.50

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

BEST COPY AVAILABLE

Table 5-11.—Reading Item Response Theory scale score (range of possible values: 0 to 92)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	22.67	8.58	18,937	32.47	10.85	5,053	37.97	12.67	16,336	54.77	14.17
Male	8,984	22.09	8.64	9,688	31.57	10.86	2,556	36.82	12.61	8,349	53.41	14.59
Female	8,640	23.29	8.47	9,247	33.44	10.76	2,497	39.18	12.62	7,987	56.22	13.56
White, non-Hispanic	10,433	23.89	8.71	11,073	34.05	10.85	2,935	39.95	12.76	9,435	57.41	13.53
Black, non-Hispanic	2,854	20.55	7.08	2,968	29.36	9.75	782	34.67	11.09	2,371	49.88	14.10
Hispanic, race specified	1,182	20.79	8.11	1,315	30.95	10.27	322	36.53	11.34	1,233	52.40	13.81
Hispanic, race not specified	1,195	19.33	6.81	1,423	28.80	9.60	377	32.46	10.83	1,335	49.15	13.49
Asian	897	26.72	11.74	1,089	37.44	13.25	257	44.02	16.42	1,042	59.09	14.44
Hawaiian or other Pacific Islander	186	21.82	8.60	202	30.34	10.01	93	32.78	10.74	188	52.69	12.97
American Indian or Alaska Native	354	17.62	6.01	344	27.01	8.83	126	27.60	9.07	298	44.62	13.32
More than one race, non-Hispanic	476	22.63	9.53	473	32.12	11.37	152	37.96	12.14	397	55.49	13.96
SES: first quintile	2,594	17.96	5.44	2,917	26.68	8.13	753	30.61	9.52	2,363	46.31	13.23
SES: second quintile	3,271	20.54	6.83	3,503	30.11	9.61	925	34.67	10.85	2,796	52.42	13.52
SES: third quintile	3,470	22.03	7.28	3,686	32.15	9.60	997	38.48	11.73	3,003	55.37	12.81
SES: fourth quintile	3,650	24.22	8.42	3,909	34.72	10.38	1,019	40.71	12.00	3,173	58.14	12.49
SES: fifth quintile	3,880	27.99	10.35	4,152	38.67	12.16	1,159	44.87	13.76	3,642	62.58	12.64
Public school	13,737	21.94	8.13	14,579	31.60	10.41	3,809	37.17	12.36	12,998	53.84	14.07
Private school	3,888	26.65	9.77	4,358	37.24	11.92	1,042	43.68	12.66	3,279	61.07	12.86

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

BEST COPY AVAILABLE

Table 5-12.—Mathematics Item Response Theory scale score (range of possible values: 0 to 64)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	19.30	7.11	19,649	27.16	8.74	5,226	32.37	9.61	16,641	42.78	9.50
Male	9,479	19.32	7.47	10,041	27.19	9.06	2,644	32.38	10.09	8,506	42.95	9.92
Female	9,156	19.27	6.71	9,606	27.12	8.38	2,582	32.37	9.07	8,135	42.60	9.02
White, non-Hispanic	10,433	21.12	7.24	11,071	29.44	8.59	2,935	34.82	9.21	9,436	45.20	8.87
Black, non-Hispanic	2,855	16.78	5.43	2,962	23.59	7.41	781	28.78	8.86	2,371	38.15	9.22
Hispanic, race specified	1,588	16.88	6.16	1,624	24.51	8.11	389	30.42	9.04	1,354	40.72	9.55
Hispanic, race not specified	1,800	15.30	5.50	1,834	22.66	7.58	486	27.13	8.57	1,518	38.91	8.68
Asian	898	22.42	8.27	1,089	30.03	9.26	256	35.18	10.28	1,042	44.00	9.51
Hawaiian or other Pacific Islander	187	17.91	6.22	202	24.55	7.61	93	28.26	7.57	188	38.97	8.25
American Indian or Alaska Native	354	15.61	5.74	345	23.35	7.53	126	25.24	8.72	298	38.04	8.85
More than one race, non-Hispanic	473	19.23	6.96	472	26.86	8.18	151	31.52	8.80	397	42.95	9.54
SES: first quintile	3,269	14.88	4.98	3,426	21.81	6.96	895	26.17	8.58	2,572	37.36	9.26
SES: second quintile	3,429	17.51	5.86	3,607	25.42	7.86	942	30.15	8.68	2,839	40.86	9.21
SES: third quintile	3,546	19.27	6.14	3,721	27.36	7.80	1,001	32.99	8.33	3,017	43.24	8.57
SES: fourth quintile	3,676	21.08	6.76	3,921	29.41	8.14	1,023	34.72	8.26	3,178	45.30	8.19
SES: fifth quintile	3,893	24.14	7.92	4,161	32.68	8.83	1,158	38.54	9.31	3,644	48.36	7.80
Public school	14,702	18.69	6.83	15,260	26.49	8.55	3,971	31.77	9.60	13,292	42.26	9.56
Private school	3,934	22.78	7.69	4,389	30.98	8.80	1,043	36.99	8.37	3,286	46.50	7.93
English version of test	17,615	19.68	7.08	18,925	27.51	8.65	5,049	32.72	9.53	16,336	42.94	9.42
Spanish version of test	1,021	13.17	4.18	724	19.01	6.43	177	24.32	7.63	305	35.47	10.34

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

BEST COPY AVAILABLE

Table 5-13.—General knowledge Item Response Theory scale score (range of possible values: 0 to 51)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,566	22.10	7.43	18,903	26.81	7.84	5,044	30.02	7.92	16,324	34.00	7.74
Male	8,944	22.28	7.57	9,665	27.06	7.93	2,550	30.28	8.09	8,341	34.30	7.79
Female	8,621	21.91	7.28	9,236	26.54	7.73	2,494	29.75	7.72	7,983	33.67	7.67
White, non-Hispanic	10,407	24.46	7.02	11,065	29.49	7.10	2,932	32.72	6.98	9,432	36.79	6.42
Black, non-Hispanic	2,840	17.25	5.90	2,954	21.74	6.81	780	25.55	6.74	2,365	29.41	7.35
Hispanic, race specified	1,175	20.00	6.96	1,312	24.74	7.32	321	28.00	7.93	1,233	30.63	8.14
Hispanic, race not specified	1,192	18.21	6.32	1,421	22.09	6.90	376	24.37	7.25	1,335	29.01	7.32
Asian	895	19.85	7.20	1,083	23.89	7.66	256	27.20	8.44	1,041	31.55	8.64
Hawaiian or other Pacific Islander	186	18.23	6.55	202	21.83	7.37	93	23.98	7.76	188	29.54	7.48
American Indian or Alaska Native	352	17.71	6.40	344	23.04	6.92	126	22.96	7.30	298	30.43	7.80
More than one race, non-Hispanic	471	21.95	6.73	472	27.03	7.07	151	30.38	7.59	395	35.10	6.54
SES: first quintile	2,583	16.85	5.67	2,910	21.19	6.62	753	23.84	6.86	2,361	28.23	7.29
SES: second quintile	3,258	20.19	6.44	3,495	24.99	7.05	923	28.20	7.14	2,793	32.58	7.22
SES: third quintile	3,463	21.97	6.65	3,680	26.88	6.93	994	29.84	7.27	3,002	34.53	6.73
SES: fourth quintile	3,638	23.79	6.85	3,901	29.00	6.98	1,018	32.38	6.66	3,169	36.52	6.25
SES: fifth quintile	3,869	27.03	7.33	4,149	32.07	7.10	1,157	35.56	6.42	3,641	39.05	5.94
Public school	13,682	21.53	7.30	14,545	26.22	7.75	3,803	29.58	7.93	12,989	33.49	7.75
Private school	3,884	25.17	7.38	4,358	30.04	7.50	1,040	33.54	6.99	3,277	37.51	6.52

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified, Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-14.—Probability of proficiency, reading level 1: letter recognition (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	0.64	0.41	18,937	0.92	0.22	5,053	0.96	0.16	16,336	0.99	0.07
Male	8,984	0.61	0.42	9,688	0.91	0.24	2,556	0.95	0.17	8,349	0.99	0.08
Female	8,640	0.68	0.40	9,247	0.94	0.19	2,497	0.97	0.14	7,987	1.00	0.05
White, non-Hispanic	10,433	0.71	0.38	11,073	0.95	0.18	2,935	0.98	0.13	9,435	1.00	0.06
Black, non-Hispanic	2,854	0.56	0.42	2,968	0.89	0.26	782	0.95	0.18	2,371	0.99	0.08
Hispanic, race specified	1,182	0.53	0.43	1,315	0.90	0.25	322	0.98	0.12	1,233	0.99	0.08
Hispanic, race not specified	1,195	0.46	0.43	1,423	0.85	0.30	377	0.90	0.25	1,335	0.99	0.07
Asian	897	0.78	0.34	1,089	0.97	0.13	257	0.98	0.10	1,042	1.00	0.05
Hawaiian or other Pacific Islander	186	0.59	0.42	202	0.89	0.25	93	0.96	0.14	188	1.00	0.00
American Indian or Alaska Native	354	0.34	0.41	344	0.82	0.31	126	0.83	0.32	298	0.99	0.08
More than one race, non-Hispanic	476	0.62	0.42	473	0.92	0.22	152	0.95	0.18	397	0.99	0.08
SES: first quintile	2,594	0.40	0.41	2,917	0.83	0.31	753	0.90	0.25	2,363	0.98	0.11
SES: second quintile	3,271	0.56	0.41	3,503	0.90	0.25	925	0.95	0.18	2,796	0.99	0.06
SES: third quintile	3,470	0.64	0.40	3,686	0.94	0.20	997	0.97	0.13	3,003	1.00	0.05
SES: fourth quintile	3,650	0.74	0.36	3,909	0.97	0.14	1,019	0.99	0.08	3,173	1.00	0.03
SES: fifth quintile	3,880	0.84	0.30	4,152	0.98	0.11	1,159	1.00	0.05	3,642	1.00	0.01
Public school	13,737	0.61	0.42	14,579	0.92	0.23	3,809	0.96	0.16	12,998	0.99	0.07
Private school	3,888	0.82	0.32	4,358	0.97	0.14	1,042	0.99	0.07	3,279	1.00	0.03

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-15.—Probability of proficiency, reading level 2: beginning sounds (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	0.29	0.33	18,937	0.67	0.33	5,053	0.81	0.27	16,336	0.96	0.14
Male	8,984	0.27	0.32	9,688	0.64	0.35	2,556	0.78	0.30	8,349	0.95	0.16
Female	8,640	0.32	0.34	9,247	0.71	0.31	2,497	0.84	0.25	7,987	0.97	0.12
White, non-Hispanic	10,433	0.34	0.34	11,073	0.73	0.31	2,935	0.86	0.24	9,435	0.97	0.11
Black, non-Hispanic	2,854	0.21	0.28	2,968	0.56	0.35	782	0.74	0.30	2,371	0.92	0.19
Hispanic, race specified	1,182	0.23	0.31	1,315	0.63	0.35	322	0.78	0.28	1,233	0.95	0.14
Hispanic, race not specified	1,195	0.18	0.27	1,423	0.56	0.36	377	0.68	0.34	1,335	0.94	0.16
Asian	897	0.42	0.37	1,089	0.78	0.28	257	0.86	0.22	1,042	0.97	0.13
Hawaiian or other Pacific Islander	186	0.27	0.33	202	0.59	0.36	93	0.68	0.31	188	0.97	0.06
American Indian or Alaska Native	354	0.12	0.23	344	0.48	0.37	126	0.52	0.35	298	0.90	0.20
More than one race, non-Hispanic	476	0.28	0.33	473	0.66	0.33	152	0.82	0.27	397	0.96	0.16
SES: first quintile	2,594	0.11	0.20	2,917	0.48	0.35	753	0.63	0.34	2,363	0.91	0.21
SES: second quintile	3,271	0.20	0.27	3,503	0.61	0.34	925	0.75	0.30	2,796	0.95	0.15
SES: third quintile	3,470	0.27	0.31	3,686	0.69	0.32	997	0.84	0.23	3,003	0.97	0.11
SES: fourth quintile	3,650	0.35	0.34	3,909	0.76	0.29	1,019	0.88	0.20	3,173	0.98	0.08
SES: fifth quintile	3,880	0.50	0.36	4,152	0.84	0.24	1,159	0.92	0.16	3,642	0.99	0.05
Public school	13,737	0.26	0.32	14,579	0.65	0.34	3,809	0.79	0.28	12,998	0.95	0.15
Private school	3,888	0.45	0.36	4,358	0.80	0.28	1,042	0.92	0.16	3,279	0.99	0.08

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander, American Indian or Alaska Native, and more than one race specified, non-Hispanic

Table 5-16.—Probability of proficiency, reading level 3: ending sounds (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	0.16	0.26	18,937	0.48	0.35	5,053	0.65	0.32	16,336	0.91	0.20
Male	8,984	0.15	0.25	9,688	0.45	0.35	2,556	0.61	0.34	8,349	0.89	0.22
Female	8,640	0.18	0.27	9,247	0.52	0.34	2,497	0.68	0.31	7,987	0.93	0.18
White, non-Hispanic	10,433	0.19	0.27	11,073	0.54	0.33	2,935	0.71	0.29	9,435	0.94	0.16
Black, non-Hispanic	2,854	0.10	0.20	2,968	0.37	0.34	782	0.55	0.34	2,371	0.85	0.26
Hispanic, race specified	1,182	0.12	0.23	1,315	0.44	0.34	322	0.61	0.34	1,233	0.89	0.21
Hispanic, race not specified	1,195	0.09	0.18	1,423	0.37	0.33	377	0.49	0.35	1,335	0.86	0.23
Asian	897	0.26	0.33	1,089	0.60	0.34	257	0.71	0.31	1,042	0.93	0.18
Hawaiian or other Pacific Islander	186	0.15	0.25	202	0.41	0.36	93	0.47	0.34	188	0.91	0.15
American Indian or Alaska Native	354	0.06	0.15	344	0.30	0.32	126	0.32	0.31	298	0.78	0.28
More than one race, non-Hispanic	476	0.16	0.27	473	0.46	0.34	152	0.65	0.31	397	0.92	0.20
SES: first quintile	2,594	0.05	0.12	2,917	0.29	0.30	753	0.43	0.34	2,363	0.82	0.27
SES: second quintile	3,271	0.10	0.19	3,503	0.41	0.33	925	0.57	0.33	2,796	0.89	0.21
SES: third quintile	3,470	0.14	0.23	3,686	0.48	0.33	997	0.68	0.29	3,003	0.93	0.16
SES: fourth quintile	3,650	0.20	0.27	3,909	0.57	0.33	1,019	0.74	0.27	3,173	0.95	0.13
SES: fifth quintile	3,880	0.32	0.33	4,152	0.67	0.30	1,159	0.81	0.24	3,642	0.97	0.09
Public school	13,737	0.14	0.24	14,579	0.46	0.34	3,809	0.63	0.33	12,998	0.90	0.21
Private school	3,888	0.27	0.31	4,358	0.63	0.32	1,042	0.80	0.23	3,279	0.96	0.11

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-17.—Probability of proficiency, reading level 4: sight words (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	0.02	0.13	18,937	0.13	0.28	5,053	0.24	0.37	16,336	0.76	0.37
Male	8,984	0.02	0.14	9,688	0.11	0.27	2,556	0.22	0.35	8,349	0.72	0.39
Female	8,640	0.02	0.13	9,247	0.14	0.30	2,497	0.27	0.38	7,987	0.81	0.34
White, non-Hispanic	10,433	0.03	0.14	11,073	0.15	0.30	2,935	0.28	0.38	9,435	0.83	0.32
Black, non-Hispanic	2,854	0.01	0.09	2,968	0.09	0.24	782	0.17	0.32	2,371	0.66	0.42
Hispanic, race specified	1,182	0.02	0.11	1,315	0.10	0.25	322	0.22	0.34	1,233	0.71	0.40
Hispanic, race not specified	1,195	0.01	0.07	1,423	0.07	0.22	377	0.13	0.29	1,335	0.62	0.42
Asian	897	0.08	0.25	1,089	0.25	0.39	257	0.41	0.46	1,042	0.83	0.33
Hawaiian or other Pacific Islander	186	0.03	0.15	202	0.12	0.27	93	0.13	0.31	188	0.71	0.39
American Indian or Alaska Native	354	0.00	0.05	344	0.05	0.17	126	0.05	0.18	298	0.49	0.44
More than one race, non-Hispanic	476	0.03	0.16	473	0.12	0.28	152	0.26	0.38	397	0.80	0.35
SES: first quintile	2,594	0.00	0.05	2,917	0.04	0.15	753	0.09	0.23	2,363	0.56	0.43
SES: second quintile	3,271	0.01	0.09	3,503	0.08	0.23	925	0.16	0.31	2,796	0.73	0.39
SES: third quintile	3,470	0.01	0.10	3,686	0.11	0.25	997	0.24	0.36	3,003	0.80	0.34
SES: fourth quintile	3,650	0.03	0.14	3,909	0.15	0.30	1,019	0.30	0.39	3,173	0.86	0.29
SES: fifth quintile	3,880	0.06	0.21	4,152	0.25	0.37	1,159	0.42	0.42	3,642	0.90	0.25
Public school	13,737	0.02	0.12	14,579	0.11	0.26	3,809	0.22	0.35	12,998	0.75	0.38
Private school	3,888	0.05	0.19	4,358	0.23	0.36	1,042	0.39	0.41	3,279	0.88	0.27

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-18.—Probability of proficiency, reading level 5: comprehension of words (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,625	0.01	0.08	18,937	0.04	0.16	5,053	0.09	0.25	16,336	0.42	0.41
Male	8,984	0.01	0.09	9,688	0.03	0.15	2,556	0.08	0.24	8,349	0.39	0.41
Female	8,640	0.01	0.08	9,247	0.04	0.16	2,497	0.10	0.26	7,987	0.46	0.41
White, non-Hispanic	10,433	0.01	0.09	11,073	0.04	0.18	2,935	0.11	0.28	9,435	0.49	0.41
Black, non-Hispanic	2,854	0.00	0.05	2,968	0.02	0.10	782	0.05	0.17	2,371	0.31	0.38
Hispanic, race specified	1,182	0.01	0.07	1,315	0.02	0.12	322	0.06	0.20	1,233	0.36	0.39
Hispanic, race not specified	1,195	0.00	0.04	1,423	0.01	0.09	377	0.03	0.14	1,335	0.27	0.37
Asian	897	0.04	0.18	1,089	0.10	0.26	257	0.24	0.38	1,042	0.56	0.42
Hawaiian or other Pacific Islander	186	0.01	0.07	202	0.01	0.08	93	0.05	0.18	188	0.34	0.40
American Indian or Alaska Native	354	0.00	0.01	344	0.01	0.07	126	0.01	0.09	298	0.18	0.31
More than one race, non-Hispanic	476	0.02	0.12	473	0.05	0.19	152	0.08	0.22	397	0.43	0.40
SES: first quintile	2,594	0.00	0.03	2,917	0.00	0.05	753	0.02	0.10	2,363	0.20	0.32
SES: second quintile	3,271	0.00	0.05	3,503	0.02	0.12	925	0.04	0.18	2,796	0.36	0.39
SES: third quintile	3,470	0.00	0.05	3,686	0.02	0.13	997	0.08	0.24	3,003	0.43	0.40
SES: fourth quintile	3,650	0.01	0.08	3,909	0.04	0.17	1,019	0.11	0.27	3,173	0.50	0.41
SES: fifth quintile	3,880	0.03	0.14	4,152	0.09	0.25	1,159	0.18	0.34	3,642	0.63	0.40
Public school	13,737	0.01	0.07	14,579	0.03	0.14	3,809	0.08	0.23	12,998	0.40	0.41
Private school	3,888	0.02	0.12	4,358	0.07	0.22	1,042	0.15	0.31	3,279	0.60	0.41

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native; and more than one race specified, non-Hispanic.

Table 5-19.—Probability of proficiency, mathematics level 1: number and shape (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	0.92	0.18	19,649	0.99	0.07	5,226	0.99	0.04	16,641	1.00	0.02
Male	9,479	0.91	0.19	10,041	0.98	0.07	2,644	0.99	0.05	8,506	1.00	0.03
Female	9,156	0.93	0.17	9,606	0.99	0.07	2,582	1.00	0.04	8,135	1.00	0.02
White, non-Hispanic	10,433	0.95	0.13	11,071	0.99	0.05	2,935	1.00	0.03	9,436	1.00	0.02
Black, non-Hispanic	2,855	0.89	0.21	2,962	0.98	0.09	781	0.99	0.06	2,371	1.00	0.03
Hispanic, race specified	1,588	0.87	0.22	1,624	0.98	0.09	389	0.99	0.04	1,354	1.00	0.01
Hispanic, race not specified	1,800	0.82	0.26	1,834	0.97	0.11	486	0.99	0.05	1,518	1.00	0.02
Asian	898	0.96	0.11	1,089	1.00	0.03	256	1.00	0.01	1,042	1.00	0.01
Hawaiian or other Pacific Islander	187	0.90	0.20	202	0.99	0.05	93	1.00	0.00	188	1.00	0.00
American Indian or Alaska Native	354	0.82	0.26	345	0.97	0.08	126	0.97	0.11	298	1.00	0.00
More than one race, non-Hispanic	473	0.93	0.15	472	0.99	0.06	151	0.99	0.03	397	1.00	0.03
SES: first quintile	3,269	0.81	0.26	3,426	0.97	0.11	895	0.99	0.05	2,572	1.00	0.03
SES: second quintile	3,429	0.90	0.20	3,607	0.98	0.08	942	0.99	0.06	2,839	1.00	0.04
SES: third quintile	3,546	0.94	0.14	3,721	0.99	0.05	1,001	1.00	0.04	3,017	1.00	0.00
SES: fourth quintile	3,676	0.96	0.11	3,921	0.99	0.04	1,023	1.00	0.02	3,178	1.00	0.01
SES: fifth quintile	3,893	0.98	0.08	4,161	1.00	0.03	1,158	1.00	0.00	3,644	1.00	0.01
Public school	14,702	0.91	0.19	15,260	0.98	0.07	3,971	0.99	0.05	13,292	1.00	0.02
Private school	3,934	0.97	0.10	4,389	0.99	0.05	1,043	1.00	0.00	3,286	1.00	0.02
English version of test	17,615	0.93	0.17	18,925	0.99	0.06	5,049	0.99	0.04	16,336	1.00	0.02
Spanish version of test	1,021	0.73	0.30	724	0.94	0.15	177	0.99	0.03	305	1.00	0.04

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native; and more than one race specified, non-Hispanic.

Table 5-20.—Probability of proficiency, mathematics level 2: relative size, etc. (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	0.53	0.35	19,649	0.83	0.25	5,226	0.91	0.19	16,641	0.98	0.09
Male	9,479	0.53	0.36	10,041	0.82	0.26	2,644	0.90	0.20	8,506	0.98	0.10
Female	9,156	0.54	0.35	9,606	0.83	0.24	2,582	0.92	0.17	8,135	0.98	0.09
White, non-Hispanic	10,433	0.63	0.33	11,071	0.89	0.20	2,935	0.95	0.13	9,436	0.99	0.08
Black, non-Hispanic	2,855	0.41	0.33	2,962	0.74	0.29	781	0.86	0.23	2,371	0.96	0.13
Hispanic, race specified	1,588	0.41	0.34	1,624	0.75	0.29	389	0.88	0.21	1,354	0.97	0.09
Hispanic, race not specified	1,800	0.32	0.32	1,834	0.69	0.31	486	0.82	0.25	1,518	0.97	0.10
Asian	898	0.66	0.32	1,089	0.89	0.18	256	0.94	0.12	1,042	0.99	0.06
Hawaiian or other Pacific Islander	187	0.47	0.33	202	0.77	0.28	93	0.89	0.15	188	0.98	0.05
American Indian or Alaska Native	354	0.34	0.34	345	0.73	0.29	126	0.77	0.30	298	0.97	0.11
More than one race, non-Hispanic	473	0.53	0.34	472	0.84	0.23	151	0.90	0.20	397	0.98	0.10
SES: first quintile	3,269	0.30	0.30	3,426	0.67	0.31	895	0.80	0.26	2,572	0.96	0.13
SES: second quintile	3,429	0.45	0.34	3,607	0.79	0.27	942	0.90	0.20	2,839	0.97	0.11
SES: third quintile	3,546	0.55	0.33	3,721	0.86	0.21	1,001	0.94	0.15	3,017	0.98	0.08
SES: fourth quintile	3,676	0.64	0.32	3,921	0.90	0.18	1,023	0.96	0.12	3,178	0.99	0.06
SES: fifth quintile	3,893	0.75	0.28	4,161	0.94	0.14	1,158	0.97	0.08	3,644	1.00	0.03
Public school	14,702	0.51	0.35	15,260	0.81	0.26	3,971	0.90	0.20	13,292	0.98	0.10
Private school	3,934	0.70	0.30	4,389	0.91	0.17	1,043	0.98	0.07	3,286	0.99	0.05
English version of test	17,615	0.56	0.35	18,925	0.84	0.24	5,049	0.92	0.18	16,336	0.98	0.09
Spanish version of test	1,021	0.20	0.25	724	0.53	0.33	177	0.77	0.25	305	0.93	0.18

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native; and more than one race specified, non-Hispanic.

Table 5-21.—Probability of proficiency, mathematics level 3: number sequence, etc. (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	0.19	0.30	19,649	0.52	0.39	5,226	0.71	0.36	16,641	0.93	0.21
Male	9,479	0.20	0.31	10,041	0.52	0.39	2,644	0.70	0.37	8,506	0.92	0.21
Female	9,156	0.18	0.29	9,606	0.52	0.39	2,582	0.72	0.35	8,135	0.93	0.20
White, non-Hispanic	10,433	0.25	0.33	11,071	0.62	0.37	2,935	0.80	0.30	9,436	0.96	0.16
Black, non-Hispanic	2,855	0.09	0.20	2,962	0.36	0.37	781	0.59	0.39	2,371	0.86	0.28
Hispanic, race specified	1,588	0.11	0.24	1,624	0.40	0.39	389	0.66	0.38	1,354	0.90	0.25
Hispanic, race not specified	1,800	0.07	0.18	1,834	0.32	0.37	486	0.52	0.40	1,518	0.89	0.24
Asian	898	0.30	0.37	1,089	0.62	0.38	256	0.77	0.34	1,042	0.94	0.18
Hawaiian or other Pacific Islander	187	0.12	0.24	202	0.41	0.37	93	0.57	0.36	188	0.90	0.21
American Indian or Alaska Native	354	0.08	0.19	345	0.35	0.37	126	0.44	0.40	298	0.88	0.24
More than one race, non-Hispanic	473	0.18	0.29	472	0.51	0.38	151	0.72	0.35	397	0.93	0.20
SES: first quintile	3,269	0.05	0.15	3,426	0.28	0.34	895	0.47	0.40	2,572	0.85	0.28
SES: second quintile	3,429	0.12	0.23	3,607	0.45	0.38	942	0.66	0.36	2,839	0.91	0.22
SES: third quintile	3,546	0.17	0.27	3,721	0.54	0.38	1,001	0.77	0.31	3,017	0.95	0.17
SES: fourth quintile	3,676	0.24	0.32	3,921	0.63	0.36	1,023	0.82	0.28	3,178	0.97	0.14
SES: fifth quintile	3,893	0.38	0.37	4,161	0.74	0.33	1,158	0.88	0.24	3,644	0.98	0.09
Public school	14,702	0.17	0.28	15,260	0.49	0.39	3,971	0.69	0.37	13,292	0.92	0.21
Private school	3,934	0.32	0.36	4,389	0.68	0.35	1,043	0.88	0.22	3,286	0.98	0.10
English version of test	17,615	0.20	0.31	18,925	0.54	0.39	5,049	0.73	0.35	16,336	0.93	0.20
Spanish version of test	1,021	0.02	0.09	724	0.16	0.28	177	0.37	0.38	305	0.78	0.33

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-22.—Probability of proficiency, mathematics level 4: addition/subtraction (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	0.03	0.12	19,649	0.16	0.26	5,226	0.32	0.34	16,641	0.71	0.34
Male	9,479	0.04	0.13	10,041	0.16	0.27	2,644	0.33	0.35	8,506	0.70	0.35
Female	9,156	0.03	0.10	9,606	0.15	0.25	2,582	0.31	0.33	8,135	0.71	0.33
White, non-Hispanic	10,433	0.05	0.14	11,071	0.21	0.29	2,935	0.39	0.36	9,436	0.78	0.30
Black, non-Hispanic	2,855	0.01	0.05	2,962	0.07	0.16	781	0.21	0.28	2,371	0.56	0.36
Hispanic, race specified	1,588	0.01	0.07	1,624	0.10	0.20	389	0.26	0.31	1,354	0.64	0.36
Hispanic, race not specified	1,800	0.01	0.05	1,834	0.07	0.16	486	0.16	0.25	1,518	0.58	0.35
Asian	898	0.08	0.20	1,089	0.23	0.31	256	0.42	0.37	1,042	0.73	0.33
Hawaiian or other Pacific Islander	187	0.02	0.11	202	0.08	0.18	93	0.16	0.25	188	0.57	0.35
American Indian or Alaska Native	354	0.01	0.05	345	0.07	0.17	126	0.12	0.22	298	0.53	0.37
More than one race, non-Hispanic	473	0.03	0.11	472	0.14	0.24	151	0.27	0.31	397	0.70	0.34
SES: first quintile	3,269	0.00	0.04	3,426	0.05	0.13	895	0.14	0.25	2,572	0.52	0.36
SES: second quintile	3,429	0.01	0.06	3,607	0.11	0.20	942	0.23	0.29	2,839	0.65	0.35
SES: third quintile	3,546	0.02	0.08	3,721	0.14	0.23	1,001	0.32	0.32	3,017	0.73	0.31
SES: fourth quintile	3,676	0.04	0.12	3,921	0.20	0.27	1,023	0.38	0.34	3,178	0.79	0.28
SES: fifth quintile	3,893	0.09	0.20	4,161	0.31	0.33	1,158	0.53	0.37	3,644	0.87	0.23
Public school	14,702	0.03	0.10	15,260	0.14	0.24	3,971	0.30	0.34	13,292	0.69	0.35
Private school	3,934	0.07	0.18	4,389	0.25	0.31	1,043	0.46	0.35	3,286	0.83	0.26
English version of test	17,615	0.03	0.12	18,925	0.16	0.26	5,049	0.33	0.34	16,336	0.71	0.34
Spanish version of test	1,021	0.00	0.02	724	0.03	0.11	177	0.10	0.20	305	0.46	0.40

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic.

Table 5-23.—Probability of proficiency, mathematics level 5: multiplication/division (range of possible values: 0.0 to 1.0)*

Characteristic	Round 1			Round 2			Round 3			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,636	0.00	0.03	19,649	0.01	0.08	5,226	0.05	0.16	16,641	0.24	0.33
Male	9,479	0.00	0.04	10,041	0.02	0.10	2,644	0.05	0.17	8,506	0.26	0.34
Female	9,156	0.00	0.02	9,606	0.01	0.06	2,582	0.04	0.14	8,135	0.21	0.31
White, non-Hispanic	10,433	0.00	0.04	11,071	0.02	0.10	2,935	0.07	0.19	9,436	0.31	0.35
Black, non-Hispanic	2,855	0.00	0.02	2,962	0.00	0.04	781	0.01	0.07	2,371	0.09	0.20
Hispanic, race specified	1,588	0.00	0.02	1,624	0.01	0.05	389	0.02	0.10	1,354	0.17	0.28
Hispanic, race not specified	1,800	0.00	0.00	1,834	0.00	0.03	486	0.01	0.03	1,518	0.10	0.21
Asian	898	0.01	0.06	1,089	0.03	0.14	256	0.10	0.24	1,042	0.28	0.35
Hawaiian or other Pacific Islander	187	0.00	0.00	202	0.01	0.07	93	0.01	0.05	188	0.10	0.21
American Indian or Alaska Native	354	0.00	0.00	345	0.00	0.05	126	0.01	0.05	298	0.09	0.20
More than one race, non-Hispanic	473	0.00	0.04	472	0.01	0.07	151	0.03	0.13	397	0.25	0.33
SES: first quintile	3,269	0.00	0.01	3,426	0.00	0.02	895	0.01	0.04	2,572	0.08	0.19
SES: second quintile	3,429	0.00	0.00	3,607	0.01	0.04	942	0.02	0.11	2,839	0.16	0.27
SES: third quintile	3,546	0.00	0.01	3,721	0.01	0.06	1,001	0.03	0.12	3,017	0.22	0.31
SES: fourth quintile	3,676	0.00	0.03	3,921	0.02	0.09	1,023	0.05	0.15	3,178	0.30	0.34
SES: fifth quintile	3,893	0.01	0.07	4,161	0.04	0.15	1,158	0.13	0.26	3,644	0.44	0.38
Public school	14,702	0.00	0.03	15,260	0.01	0.07	3,971	0.04	0.15	13,292	0.22	0.32
Private school	3,934	0.01	0.05	4,389	0.03	0.13	1,043	0.09	0.22	3,286	0.35	0.36
English version of test	17,615	0.00	0.03	18,925	0.02	0.09	5,049	0.05	0.16	16,336	0.24	0.33
Spanish version of test	1,021	0.00	0.00	724	0.00	0.03	177	0.00	0.03	305	0.07	0.16

* Due to missing information on some of the variables (e.g., race/ethnicity, socioeconomic status), column numbers may not add to sample total. Respondents were asked if they were Hispanic or not. Using the six race dichotomous variables (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander) and the Hispanic ethnicity variable, the race/ethnicity composite variables were created. The categories were as follows: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified, Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaska Native, and more than one race specified, non-Hispanic

BEST COPY AVAILABLE

5.9 Interviewer Variance as a Threat to Validity

There has been some concern expressed about the individual mode of administration and how it may have contributed unwanted sources of variance to the children's performance in the direct cognitive measures. Unlike group administrations, which in theory are more easily standardized, administering assessments on an individual basis to a national sample could lead to sources of variance unique to the individual administrators that in turn might affect the between individual and/or between school components of variance. A three-level multilevel analysis (Goldstein 1995, Bryk & Raudenbush 1992) was carried out in an effort to shed some light on this possibility. Tables 5-24 and 5-25 present maximum likelihood estimates of the components of variance for team leader (level 3), test administrator or interviewer (level 2), and child (level 1) for fall-kindergarten, the point of entry to school (table 5-24) and spring-first grade (table 5-25). That is, the child is nested under interviewer, and interviewer is nested under team leader.

Table 5-24.—Components of variance associated with the child, interviewer, and team leader for fall-kindergarten

Cognitive tests	Child (level 1)	Interviewer (level 2)	Team leader (level 3)
Reading	.72 (92.3%)	.01 (1.3%)	.05 (6.4%)
Mathematics	.69 (92.0%)	.01 (1.3%)	.05 (6.6%)
General Knowledge	.70 (86.4%)	.02 (2.5%)	.09 (11.1%)

Table 5-25.—Components of variance associated with the child, interviewer, and team leader for spring-first grade

Cognitive tests	Child (level 1)	Interviewer (level 2)	Team leader (level 3)
Reading	.38 (92.7%)	.01 (2.4%)	.02 (4.8%)
Mathematics	.59 (92.0%)	.01 (1.5%)	.04 (6.3%)
General Knowledge	.28 (84.8%)	.01 (3.0%)	.04 (12.1%)

Inspection of tables 5-24 and 5-25 suggest that the interviewer source of variance is relatively trivial and ranges from a low of 1.3 percent in reading and mathematics in fall-kindergarten to a high of 3 percent on the general knowledge test in spring-first grade. While relatively trivial in terms of percentage, the interviewer effect for general knowledge tends to be about twice that of reading and mathematics. The large number of open-ended items in the general knowledge test tended to be much more subjective in nature than those in the reading and mathematics tests, and thus more vulnerable to variations in interviewers' evaluations of responses. Unfortunately, it is more difficult to interpret the source of variance associated with the team leader since team leaders tended to be associated with primary sampling units. It is interesting to note the reduction in absolute terms of the between child variance as one moves from fall-kindergarten to spring-first grade. In general, there was also a proportional reduction in between school variance on the theta scale as the children move through the school system.

6. PSYCHOMETRIC CHARACTERISTICS OF THE INDIRECT AND PSYCHOMOTOR MEASURES

Chapter 2 describes the selection and development of the indirect and psychomotor measures. This chapter provides details of their psychometric characteristics in the kindergarten and first-grade rounds of data collection. In addition, the relationships between the direct and indirect cognitive measures are explored.

In the fall- and spring-kindergarten and spring-first-grade data collections (rounds 1, 2, and 4), teachers of the sampled children were asked to evaluate each child's academic and social skills. Parents also rated social skills. These measures were not collected in fall-first grade, round 3, when a subsample of children was tested on the direct cognitive measures only. The psychomotor test, measuring children's fine and gross motor skills, was administered in round 1 only, at entry to kindergarten. Appendix E presents score statistics on each of these measures for selected subgroups. Additional details may be found in the Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K) user manuals.

Differential item functioning (DIF) analysis was not carried out for the indirect and psychomotor measures. DIF assumptions are not really relevant to behavioral, physical, and attitudinal measures. The idea of DIF is that for subsets of individuals *matched on ability level* (based on the total set of items or some external criterion) *similar item performance* for different subgroups should be observed. Significant deviation from this could indicate that an item is measuring differently for different groups. For behavioral measures such as the Social Rating Scale (SRS), there is no expectation that ratings would be the same for different groups. Any group differences in ratings may reflect either legitimate real differences in the group's attitude or behavior on an item or set of items, or factors having to do with the standards or attitudes of the rater (parent or teacher), not differential functioning or flaws in the item.

DIF analysis of the Academic Rating Scale (ARS) was not appropriate for several reasons. First, the teacher produced the ratings, not by direct observation of the child. Therefore, there is an additional confounding source of difference, namely the teacher's attitudes or potential bias that cannot be separated from the child's performance. Second, even if it could be determined that teacher ratings were completely accurate and unbiased, DIF would also be impossible for the ARS because there is no satisfactory criterion for matching. The scales are too short (i.e., each item represents too big a part of the total score needed for matching), and there is no independent measure of the same construct. The direct cognitive score would not be an appropriate criterion because the ARS covers process questions that are not represented in the direct cognitive tests. Third, factor analysis of the ARS scales found a very strong first factor, which suggests that a "halo" effect is operating. This suggests that DIF analysis using total ARS score as the criterion would probably find no evidence of DIF simply because a teacher who rates a child high on one item tends to rate the same child high on all items. It is probably not *items* that are functioning differently here, but it may be *teachers* differentially rating children. This is not a psychometric characteristic of the scale itself. It is possible that the interaction between parents' and teachers' attitudes and demographic characteristics, and the demographic characteristics, cognitive ability, and behavior of children may influence the social and academic ratings assigned to children. Secondary analysis of these relationships may reveal differences in the standards used in the SRS and ARS ratings.

The psychomotor assessment consisted of five fine motor and four gross motor tasks. These scales were not long enough to provide an internal criterion score for DIF, and no external criterion was available. There also could be no assumption that fine and gross motor tasks should be measuring the same construct.

6.1 Indirect Cognitive Assessment Using the Academic Rating Scale

Teachers of ECLS–K students rated the children's academic achievement at three points in time, fall- and spring-kindergarten (rounds 1 and 2) and spring-first-grade (round 4). The ARS evaluated achievement in the three domains that are also directly assessed in the cognitive battery: language and literacy (reading), general knowledge (science and social studies), and mathematical thinking. The ARS was designed both to overlap and to augment the information gathered through the direct cognitive assessment battery. Although the direct and indirect instruments measure children's skills and behaviors within the same broad curricular domains with some intended overlap, several of the constructs they were designed to measure differ in significant ways. Most importantly, the ARS includes items designed to measure both the process and products of children's learning in school, whereas the direct cognitive battery assesses only the products of children's achievement. The scope of curricular content represented in the indirect measures is designed to be broader than the content represented on the direct cognitive measures. Unlike the direct cognitive measures, which were designed to measure gain on a longitudinal vertical scale from kindergarten entry through the end of first grade, the ARS is targeted to a specific grade level. The questions range from items with explicitly objective elements (e.g., "names all upper-

and lower-case letters of the alphabet") to others with a more subjective element (e.g., "composes simple stories" or "uses a variety of strategies to solve mathematics problems"). Teachers evaluating the children's skills were instructed to rate each child compared with other children of the same age level. Response options for each item ranged from 1 ("not yet") to 5 ("proficient"). See chapter 2, section 2.3 for additional details on the design and development of the ARS instrument.

6.2 Item Response Theory

6.2.1 One Parameter Item Response Theory

A Rasch model (Rasch 1960) was used to estimate the scores on the ARS. In Rasch models (also called a one-parameter logistic models), the log odds of the probability of a correct response is a function of the difference between the person's ability and the difficulty of the item. The item discrimination is held constant across the items, and there is no guessing parameter. Applying the Rasch model to the data allows one to construct invariant linear measures, estimate the accuracy of the measures (standard errors), and determine the degree to which these measures and their errors are confirmed in the data using the fit statistics (Wright 1999). Like the three parameter IRT models, the Rasch model assumes unidimensionality, that is, a single dimension is being measured.

The Rasch Rating Scale model (Wright & Masters 1982) was used with the ARS data:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_k)]}{\sum_{k=o}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_k)]}, \quad x = 0.1. \ldots ,m \tag{6.1}$$

where $\tau_0 = 0$ so that $\exp \sum_{j=0}^{0} \left[\beta_n - \left(\delta_i + \tau_j\right)\right] = 1$

π_{nix} is the probability that for child *n* the teacher chooses category x of ARS item *i*,

β_n is a person measure indicating the location of child *n* on the variable (e.g., Mathematical Thinking) being measured,

δ_i is the "difficulty" of ARS item *i*,

τ_k are response thresholds, or "step difficulties" for each response category on the rating scale, and

m is the maximum category number,

x is the current category, and j and k are suffixes that vary between 0 and m.

An easier to understand derivation of this model (Wright, 1999) is:

$$\text{Log}(\pi_{nix}/\pi_{ni(x-1)}) = \beta_n - \delta_i - F_x \tag{6.2}$$

where,

π_{nix} is the probability that for child n the teacher chooses category x on ARS item i,

β_n is a person measure indicating the location of person n on the variable (e.g., Mathematical Thinking)) being measured,

δ_i is the difficulty of ARS item i, and

F_x are response thresholds, or "step difficulties" for each response category in the rating scale.

β_n is comparable to the theta described in the chapter on the three parameter IRT model used in estimating the scores for the direct measures.

6.2.2 Item Response Theory Estimation Using Winsteps

Winsteps software (Linacre & Wright 2000) uses joint maximum likelihood estimation. PROX is used for the initial estimates and then UCON is used, for the final iterations. PROX assumes a normal distribution and does not take advantage of the ability of Rasch to calibrate measures independent of the sample characteristics (Wright & Masters 1982) but provides a good starting point for the estimates. UCON does not assume a normal distribution and performs a simultaneous estimation of the person and item parameters. With Winsteps, UCON is adjusted for the bias based on the length of the test (L/(L-1)) (Wright & Masters 1982). Maximum scores are excluded for calibration of the items. Winsteps provides a variety of fit statistics and a factor analysis of the residuals.

Reliability estimates are provided for both the item and persons and indicate the replicability of the placement of the persons and items. The person reliability is analogous to Cronbach alpha. Fit statistics are also provided for both persons and items (table 6-1). Both an information-weighted (infit) and an outlier sensitive (outfit) statistic are provided. The outfit mean square is sensitive to unexpected response on items far from the person's trait level. The infit mean square is weighted for the variance of the residual and thus is more influenced by unexpected responses close to the person's trait level (Linacre & Wright 2000). The expected value for the mean square is 1.0. For samples larger than 1000, fit statistics greater than 1.1 indicate departures from expected response patterns that should be examined (Smith, Schumacker, & Bush 1995).

The reliability for each of the scales was very high. The summary fit statistics for items and persons were acceptable for all the scales (see table 6-2). The fit statistics for the step calibrations indicate that the lowest category ("Not yet") was used less than expected.

Table 6-1.—Person reliability for the Rasch-based score

Category	Spring-kindergarten	Spring-first grade
ARS Language and Literacy	.91	.94
ARS Mathematical Thinking	.94	.94
ARS General Knowledge	.95	.95

Table 6-2.—Fit statistics for Persons and Items

Category	Infit MNSQ	Outfit MNSQ
Kindergarten		
Persons		
Language and Literacy	1.09	1.10
Mathematical Thinking	.95	.95
General Knowledge	1.00	1.00
Items		
Language and Literacy	1.07	1.07
Mathematical Thinking	1.01	.97
General Knowledge	1.04	1.04
Grade 1		
Persons		
Language and Literacy	1.04	1.02
Mathematical Thinking	1.05	1.05
General Knowledge	.94	.94
Items		
Language and Literacy	1.04	1.02
Mathematical Thinking	1.06	1.05
General Knowledge	.98	.94

The ARS scores were scaled to have a low of one and a high of five to correspond to the five-point rating scale that teachers used in rating children on these items but should not be interpreted as mean scores. The item difficulties and student scores are placed on a common scale. Students have a high probability of receiving a high rating on items below their scale score and a lower probability of receiving a high rating on items above their scale score. For example, a child whose Rasch IRT scale score is 4.0 would have a greater than 50 percent probability of having received a rating of "5" on all items whose difficulty is below 4.0 on the scale. Students who received maximum ratings on all the items or minimum ratings on all the items are assigned an estimated score using Bayesian techniques.

Different sets of item ratings were developed for the fall-kindergarten, spring-kindergarten, and spring-first-grade ARS instruments. Although the item stems are similar across grades, the extended item descriptions include performance criteria that increase from one grade to the next. There was sufficient overlap of identical items in the fall- and spring-kindergarten forms that a common calibration could be carried out. Because the metric is the same, change scores may be computed for the kindergarten year. The first-grade items differ from the kindergarten items and are placed on a different metric. Change scores should *not* be used to compare ratings on the first-grade scale with kindergarten performance.

Although the Rasch analysis can estimate a score based on the responses given even when there is missing data, scores estimated on a limited number of responses are less reliable than scores with more ratings. Scores were included on the data file only if at least 60 percent of the items were given ratings. The weighted means and standard deviations for the ARS scores in rounds 1, 2, and 4 are shown in table 6-3. The ARS was not administered in round 3, fall-first grade. Kindergarten repeaters were rated on the kindergarten ARS scale in round 4 rather than the first-grade form. Their scores are excluded from the means shown in the tables; only first graders are reported here.

Table 6-3.—Academic rating scale means and standard deviations (range of possible values: 1 to 5)

Category	Round 1	Round 2	Round 4
Language and Literacy	2.48 (0.73)	3.33 (0.81)	3.40 (0.93)
Mathematical Thinking	2.54 (0.82)	3.50 (0.86)	3.43 (0.90)
General Knowledge	2.62 (0.98)	3.55 (0.99)	3.26 (0.99)

6.2.3 Floor and Ceiling

As noted in the section on the development of the ARS, the criteria for some of the items was set very high to avoid serious ceiling problems and some items were included at a level designed to avoid most floor problems. Because teachers would not respond to items far outside the range of grade-level performance (they would have little opportunity to observe this as well), it is unavoidable in this type of measure that some children will have perfect scores. Table 6-4 presents the percentage of children at the ceiling and floor of the measures.

The items on each of the measures tended to cluster in difficulty particularly on the general knowledge scale. Tables showing item difficulties are provided for both kindergarten and first grade items in the ECLS–K user manuals.

Table 6-4.—Percent of sample with perfect and minimum academic rating scale scores in kindergarten and first grade

Category	Round 1	Round 2	Round 4
Percent perfect scores			
Language and Literacy	1.1	7.1	9.0
Mathematical Thinking	1.3	8.5	9.1
General Knowledge	2.9	14.2	8.8
Percent minimum scores			
Language and Literacy	6.4	0.9	1.2
Mathematical Thinking	4.6	0.7	1.3
General Knowledge	5.4	1.0	1.7

6.3 Social Rating Scale

The SRS is an adaptation of the Social Skills Rating System (Gresham & Elliott 1990). Both teachers and parents used a frequency scale (see table 6-5) to report on how often the student demonstrated the social skill or behavior described (1=never to 4=very often). Factor analyses (both exploratory analyses and confirmatory factor analyses using LISREL) were used to confirm the scales. The scale scores on all SRS scales are the mean of the ratings on the items included in the scale. Scores were computed only if the student was rated on at least two-thirds of the items in that scale. The same items were administered in fall- and spring-kindergarten. The reliability for the teacher SRS scales is high (see table 6-5). The reliability is considerably lower for the parent scales (see table 6-6).

Table 6-5.—Teacher social rating scales: split half reliability

Characteristic	Round 1	Round 2	Round 4
Approaches to learning	0.89	0.89	0.89
Self-control	0.79	0.80	0.80
Interpersonal	0.89	0.89	0.89
Externalizing problem behaviors	0.90	0.90	0.86
Internalizing problem behaviors	0.80	0.78	0.77

Table 6-6.—Parent social rating scales: split half reliability

Characteristic	Round 1	Round 2	Round 4
Approaches to learning	0.68	0.69	0.69
Self-control	0.74	0.75	0.75
Social interaction	0.70	0.68	0.69
Impulsive/overactive	0.46	0.47	0.48
Sad/lonely	0.60	0.61	0.63

Care should be taken when entering these scales into the same analysis due to problems of multicollinearity. The intercorrelations among the five teacher SRS factors are generally high. The correlations among the teacher SRS factors range from about .25 to .80 in all rounds, with the correlations among approaches to learning, self-control, and interpersonal consistently at about .65 or higher. Only the internalizing problem behaviors scale had substantially weaker relationships with the other measures, with correlations generally in the .30s or lower. Means and standard deviations for the teacher SRS are presented in table 6-7.

The items on the parent SRS were administered as part of a longer telephone or in-person survey. The factors on the parent SRS are similar to the teacher SRS; however, the items in the parent SRS are designed for the home environment and, thus, are not the same items. It is also important to keep in mind that parents and teachers observe the children in very different environments. The correlations among the parent SRS factors were not as high as for the teacher scales. Correlations between approaches to learning and self-control were consistently in the mid- .40s, while self-control correlated in the -.40s with impulsive/overactive. Most other intercorrelations among the parent scales were in the .20s or lower. Means and standard deviations for the parent SRS scales are shown in table 6-8.

Table 6-7.—Teacher social rating scales: means and standard deviations (range of possible values: 1 to 4)

Characteristic	Round 1	Round 2	Round 4
Approaches to learning	3.0 (0.7)	3.1 (0.7)	3.0 (0.7)
Self-control	3.1 (0.6)	3.2 (0.6)	3.2 (0.6)
Interpersonal	3.0 (0.6)	3.1 (0.6)	3.1 (0.6)
Externalizing problem behaviors	1.6 (0.6)	1.7 (0.7)	1.7 (0.6)
Internalizing problem behaviors	1.6 (0.5)	1.6 (0.5)	1.6 (0.5)

Table 6-8.—Parent social rating scales: means and standard deviations (range of possible values: 1to 4)

Characteristic	Round 1	Round 2	Round 4
Approaches to learning	3.1 (0.5)	3.1 (0.5)	3.1 (0.5)
Self-control	2.8 (0.5)	2.9 (0.5)	3.0 (0.5)
Social interaction	3.3 (0.6)	3.4 (0.5)	3.4 (0.5)
Sad/lonely	1.5 (0.4)	1.6 (0.4)	1.5 (0.4)
Impulsive/overactive	2.0 (0.7)	2.0 (0.7)	1.9 (0.7)

6.4 Psychomotor Assessment

The psychomotor assessment includes two scales, one measuring fine motor skills (eye-hand coordination) and the other measuring gross motor skills (balance and motor planning). The psychomotor test was administered only once, at entry to kindergarten. The internal consistency of the scales was constrained by the limited number of items in each scale combined with the diversity of motor skills measured and the limited variance in item scores (maximum score on items was 1 to 2). Alpha coefficients (reliabilities) were 0.57 for fine motor skills, 0.51 for gross motor skills, and 0.61 for the composite motor skills. Means and standard deviations for the three scales are shown in table 6-9.

Table 6-9.—Psychomotor scales: means and standard deviations (round 1 only)

Fine motor skills (0-9)	5.7 (2.1)
Gross motor skills (0-8)	6.3 (1.9)
Composite motor skills (0-17)	12.1 (3.1)

6.5 Discriminant and Convergent Validity of the Direct and Indirect Measures

As indicated earlier, the patterns of correlations among selected measures provide evidence for their construct validity, that is, whether they measure what they purport to measure. Systematic evidence for construct validity is often described in terms of *convergent* and *discriminant* validity. Convergent validity means that two different measures of the *same* trait or skill ought to have relatively high correlations with each other. Conversely, discriminant validity means that two measures that are designed to measure two *different* traits or skills should show lower correlations with each other than each does with its matching measure. (An exception to this model is high correlations that may be found for different measures that constitute a cause and effect.) More complete discussions of construct validity may be found in Campbell & Fiske (1959) and Campbell (1960).

Ten measures were intercorrelated within rounds 1, 2, and 4. These measures included the three direct and three indirect cognitive test scores and subsets of the teacher and parent social skills ratings. These correlations are shown in table 6-10 by round. Round 3 was not included because the indirect measures were not collected in fall-first grade. The ten measures were as follows:

1.	ArsLit	teacher ARS score for Language and Literacy
2.	ArsMth	teacher ARS score for Mathematical Thinking
3.	ArsGnk	teacher ARS score for General Knowledge
4.	AppLrnT	teacher SRS factor score for Approaches to Learning
5.	IntPersT	teacher SRS factor score for Social Interaction
6.	SelfConT	teacher SRS factor score for Self-Control
7.	AppLrnP	parent SRS factor score for Approaches to Learning
8.	Reading	direct cognitive test theta (ability) estimate for Reading
9.	Mathematics	direct cognitive test theta (ability) estimate for Mathematics
10.	GenK	direct cognitive test theta (ability) estimate for General Knowledge

Table 6-10.—Intercorrelations among the indirect cognitive teacher ratings, selected teacher and parent socio-behavioral measures, and direct cognitive test scores

Round 1

Measures	ArsLit	ArsMth	ArsGnk	AppLrnT	IntpersT	SelfConT	AppLrnP	Reading	Mathematics	GenK
ArsLit	1.00									
ArsMth	.82	1.00								
ArsGnk	.80	.86	1.00							
AppLrnT	.54	.51	.48	1.00						
IntpersT	.37	.37	.37	.71	1.00					
SelfConT	.26	.29	.28	.67	.78	1.00				
AppLrnP	.22	.19	.17	.22	.16	.14	1.00			
Reading	.62	.53	.46	.42	.26	.22	.20	1.00		
Mathematics	.61	.55	.49	.45	.28	.22	.24	.77	1.00	
GenK	.49	.46	.43	.37	.26	.23	.20	.57	.64	1.00

Round 2

Measures	ArsLit	ArsMth	ArsGnk	AppLrnT	IntpersT	SelfConT	AppLrnP	Reading	Mathematics	GenK
ArsLit	1.00									
ArsMth	.84	1.00								
ArsGnk	.80	.86	1.00							
AppLrnT	.60	.59	.54	1.00						
IntpersT	.39	.40	.40	.70	1.00					
SelfConT	.31	.33	.31	.66	.80	1.00				
AppLrnP	.24	.22	.21	.26	.18	.15	1.00			
Reading	.69	.59	.51	.47	.27	.24	.22	1.00		
Mathematics	.63	.60	.53	.47	.27	.24	.23	.76	1.00	
GenK	.47	.47	.45	.35	.24	.21	.20	.57	.66	1.00

Round 4

Measures	ArsLit	ArsMth	ArsGnk	AppLrnT	IntpersT	SelfConT	AppLrnP	Reading	Mathematics	GenK
ArsLit	1.00									
ArsMth	.82	1.00								
ArsGnk	.80	.83	1.00							
AppLrnT	.63	.58	.53	1.00						
IntpersT	.38	.35	.35	.69	1.00					
SelfConT	.30	.30	.28	.64	.80	1.00				
AppLrnP	.24	.20	.18	.26	.18	.15	1.00			
Reading	.72	.61	.55	.47	.26	.23	.24	1.00		
Mathematics	.58	.61	.51	.45	.25	.22	.24	.74	1.00	
GenK	.47	.47	.44	.33	.22	.19	.20	.59	.67	1.00

Indirect ARS measures 1 to 3 have counterparts in measures 8 to 10, the direct cognitive test scores. It is instructive to compare the discriminant validity of the two sets of cognitive measures (the extent to which scores measuring different constructs should be different) and the convergent validity (the extent to which scores should be closely related to other measures of the same construct). The correlation of the ARS language/literacy measure with ARS mathematical thinking varies from .82 to .84 in the three rounds of kindergarten and first grade. The comparable correlations for the direct cognitive measures of reading and mathematics range from .74 to a high of .77. It is also interesting to note that the direct measures of reading and mathematics show a slight decrease in their correlation from round 1 to round 4, suggesting the possibility of some divergence of the two skills. The correlations of the ARS general knowledge scale with the language/literacy and mathematical thinking measures are also consistently high (.80 to .86), while the direct cognitive correlations for general knowledge are substantially lower (.57 to .59 with reading, and .64 to .67 with mathematics). The differences between the two sets of correlations suggests somewhat less discriminant validity for the ARS than for the direct measures.

When one examines the cross correlations from a convergent validity perspective, differences between the indirect and direct measures are also found. One would expect that the ARS score for language/literacy would be more closely related to the direct measure of reading than to the direct measures of mathematics and general knowledge. This was true for rounds 2 and 4, but in round 1, fall-kindergarten, the ARS language and literacy scale has almost identical correlations with the reading and mathematics direct measures. The evidence for convergent validity of the ARS mathematics measure was more problematic: in all three rounds, correlations of the ARS mathematical thinking score with the direct cognitive reading were almost exactly the same as those with the direct mathematics score. The ARS general knowledge correlations were substantially *higher* with both direct mathematics and direct reading scores than they were with the direct measure of general knowledge that should have been a closer match. This pattern for general knowledge was consistent for all three rounds, with the gap increasing over time. The finding of relatively lower convergent validity for the indirect cognitive measures is a consequence of the consistently high correlations among the measures (.80s). Correlations this high mean that the measures are unlikely to show strong differential relationships with other external measures, even if those external measures are designed to assess similar constructs.

The indirect cognitive measures also show consistently higher relationships with behavioral scales such as teacher ratings of approaches to learning, interpersonal behavior, and self control than do the comparable direct cognitive measures (table 6-10). The higher intercorrelations among the indirect cognitive measures may be partly due to the fact that they do indeed measure process in addition to products. Teachers' views of children's attitudes and behavior could influence their ratings of all content domains, as well as the other socio-behavioral measures. However, regardless of the reason(s) for the greater "halo" effect, one is less likely to find differential relationships with other external process or skill measures. An additional consequence of having a significant part of the "halo" effect coming from the sharing of the learning process variable "approaches to learning" is that the indirect cognitive scale scores are somewhat more difficult to interpret. Johnny could have the same high score as Jennifer but Johnny got his score by being high on approaches to learning and low on the skill, while Jennifer was low on approaches to learning but high on the skill/knowledge purported to being assessed.

7. PSYCHOMETRIC CHARACTERISTICS OF THE ENGLISH AND SPANISH ORAL LANGUAGE DEVELOPMENT SCALE

Prior to administration of the direct cognitive assessments, the Oral Language Development Scale (OLDS), a selection of three subtests of the PreLAS 2000 (Duncan and DeAvila, 1998), was given to children who were identified by school records or teachers as having a non-English language background. Children who scored 37 or above (out of a possible 60 points) were administered the cognitive assessment in English and were not retested for English proficiency in subsequent rounds of data collection. See section 2.6 for more details of the instrument used for measuring English proficiency.

At kindergarten entry, about 15 percent of the ECLS–K participants were found to need screening for English proficiency, and about half of those screened demonstrated sufficient English language skills to participate in the cognitive assessments in English. By round 4, spring of first grade, less than 6 percent of the sample were screened, and nearly two-thirds of them achieved the score of 37 or higher required to go on to the rest of the assessment. The numbers reported in the following tables (7-1 and 7-2) are unweighted and describe patterns observed in the ECLS–K sample. They do not purport to be representative of the national population. Note that numbers in the tables may differ slightly from the public use files because a few participants were excluded from the final released samples.

In the first round of testing, about 10 percent of the screened children (16 percent of Spanish-speaking children) were so unfamiliar with English that they received zero scores on the OLDS measure. This changed dramatically by the end of kindergarten, with only 3 percent of children (4 percent of Spanish-speaking children) with zero scores. By the end of first grade, 99 percent of the screened children were able to respond to at least some of the OLDS questions, with more than 80 percent scoring 25 or above, and nearly two-thirds passing the criterion score of 37.

Table 7-1.—English Oral Language Development Scale test results

Category	Round 1	Round 2	Round 3*	Round 4
Total ECLS–K sample	19,162	19,927	5,269	16,690
Total OLDS sample	2,865	1,654	523	945
Spanish speakers	1,770	1,141	375	696
Other languages	1,095	513	148	249
Percent passing (37+)				
Total OLDS sample	49%	42%	59%	63%
Spanish speakers	41%	36%	52%	56%
Other languages	61%	55%	78%	84%
Mean (s.d.) OLDS score				
Total OLDS sample	30.5 (18.3)	29.9 (15.9)	37.7 (17.6)	38.5 (14.7)
Spanish speakers	26.7 (19.2)	27.3 (16.6)	34.9 (18.4)	36.2 (15.1)
Other languages	36.8 (14.6)	35.7 (12.6)	44.8 (12.7)	45.0 (11.0)
Percent zero scores				
Total OLDS sample	10%	3%	1%	1%
Spanish speakers	16%	4%	1%	1%
Other languages	1%	0%	0%	0%
Split-half reliability	.97	.96	.98	.96

* Fall-first grade is a subset of approximately 30 percent of the full ECLS–K sample.

Results differed for Spanish-speaking children compared with children of other language groups. At each round of testing, mean OLDS scores for Spanish speakers were more than half a standard deviation lower than those of the other language speakers. Children in the other language groups were about one and one-half times more likely to achieve a passing score on the OLDS measure at each round than were Spanish speakers. At entry to kindergarten, 41 percent of the children from Spanish-speaking backgrounds who were screened possessed the skills necessary to take the ECLS–K assessment battery, compared with 61 percent of children from other language minority groups. By spring of first grade, 307 Spanish-speaking children (44 percent of those tested in round 4), and 40 children from other language groups (16 percent), still did not achieve the cutting score necessary to participate in the ECLS–K assessments in English.

Split-half reliability coefficients were extremely high for the English OLDS test, .96 or above at all rounds. The high reliabilities are due in part to the weight assigned to the "Let's Tell Stories" part of the test, which accounted for 40 of the 60 possible score points. Each of two stories was scored on a 0 to 5 scale, and the individual story scores were multiplied by 4. The stories were of about the same difficulty, with most of the children receiving the same score on the two stories, or scores that differed by only one point. However, alpha coefficients for the Simon Says and Art Show subtests were also high (mostly mid .80s to mid .90s for both subtests, both language groups and total, and all four rounds: very high for subtests with only ten items each), and intercorrelations among the three subtests were high (.58 to .84), so very high reliability coefficients would have been obtained even without the disproportionate weight on the story scores.

Spanish-speaking children who failed to achieve a score of 37 on the English OLDS test were administered the Spanish-language version as a measure of their knowledge of Spanish. Table 7-2 presents results for this test administration.

Table 7-2.—Spanish Oral Language Development Scale test results

Category	Round 1	Round 2	Round 3*	Round 4
Number tested	1,039	728	180	307
Mean (s.d.) score	38.9 (11.9)	42.0 (9.9)	23.6 (7.4)	24.9 (6.1)
Split-half reliability	.92	.91	.92	.91

* Fall-first grade is a subset of approximately 30 percent of the full ECLS–K sample.

The decline in mean scores between kindergarten and first grade does not indicate a decline in Spanish-speaking ability but reflects the different sample of children tested each time. The Spanish OLDS test was administered only to children who failed to meet the required English OLDS cutting score. As more and more children in the later rounds passed this point, fewer and fewer were tested. (In round 3, only a subset of ECLS–K children were sampled.) The declining mean scores suggest that the Spanish-speaking children whose Spanish skills were strongest were also more likely to pass the English OLDS cutting score criterion and to leave the Spanish OLDS sample by the later rounds. The decline in standard deviations that accompanies the decline in mean scores also indicates less variation in the language ability of the Spanish-speaking children who were still taking the Spanish OLDS test in the later rounds.

Zero scores on the Spanish OLDS test are not presented in this table. There were very few such scores, and almost all were in round 1 (17 out of 1,039 tested). Probably these zero scores reflect the non-English-speaking fall kindergartners' inability to cope with the testing situation rather than lack of knowledge of Spanish. More than 80 percent of the children taking the Spanish OLDS test in fall-kindergarten achieved scores of 30 or higher. Percentages reaching a criterion score are also not presented in the table, since no performance criterion was set for participation in the Spanish translation of the ECLS–K mathematics and psychomotor assessments.

Split-half reliabilities for the Spanish OLDS test were high, for the same reasons as for the English version: consistent story scores, greater weight given to story scores, high internal consistency in the other subtests, and high correlations among subtests. The subtest intercorrelations were somewhat lower (.28 to .69) than for the English OLDS, but still high enough to support a high overall level of reliability.

Differential item functioning (DIF) analysis was not carried out for the English and Spanish OLDS tests because a satisfactory criterion score for matching equivalent groups was not available. The subtests had too few items to provide an internal criterion score for separate DIF analysis of subtests. The total score, based on the three subtests combined, was dominated by the two story ratings that together accounted for 40 of the 60 score points, making it unsuitable for a DIF analysis criterion.

8. APPROACHES TO MEASURING CHANGE USING ECLS–K LONGITUDINAL SCORES

The cognitive tests in the ECLS–K are designed to measure achievement gains over time, with the objective of relating those gains to background and educational processes. Test scores in reading, mathematics, and general knowledge are put on a common scale so that longitudinal gains can be analyzed. This chapter demonstrates a number of different analytic approaches to measuring cognitive growth that become available when one has a multiple criterion referenced developmental model. Each different analytic approach brings additional insights with respect to understanding student growth. The ECLS–K provides scale scores based on the total item pool in each of the three domains, reading, mathematics, and general knowledge. In addition, proficiency-level scores are supplied for subsets of items in reading and mathematics. (See chapter 3 and the ECLS–K users manuals for detailed descriptions of the scores.) The methodology suggested here can be carried out on as little as two longitudinal time points. The analysis described in this chapter focuses on growth in reading achievement during the kindergarten year and illustrates differences in results obtained using the total scale scores compared with the proficiency-level scores.

This example analysis focuses on (1) an individual level variable (gender) and its relationship with gains and (2) a school level variable (school sector) and its relationship with gains. In addition, this analysis examines the traditional approaches to measuring gain and shows where they may be uninformative in their conclusions about who gains and how much. It is argued that unless one explicitly takes into consideration the location of the gain on the developmental scale, the answers given by the traditional approaches may be misleading. The presence of adaptive measurement procedures makes consideration of location of gain even more important.

The sample selected to illustrate the analytic approaches to measuring change consists of ECLS–K children who had reading scores in fall- and spring-kindergarten and who stayed in the same school for the kindergarten year. In addition, the analysis sample was further restricted to children having data on parents' education (higher of father and/or mother). The final analysis sample consisted of 13,701 children in approximately 60 private non-Catholic schools, 99 Catholic schools, and almost 700 public schools. This example uses scale scores based on the total reading item pool and the reading proficiency probability scores described in chapter 3 to focus on changes taking place during the kindergarten year. Since the sampling procedure used in the ECLS–K was a multistage procedure with oversampling of certain subpopulations, all analyses reported here use a panel weight and either the survey procedures in the STATA software (2000) or multilevel approaches to correct standard errors for clustering effects.

8.1 Total Scores to Measure Longitudinal Change

The ECLS–K reports two types of scores that are based on the total item pool: thetas (standardized estimates of ability) and Item Response Theory (IRT) scale scores (estimates of number right on the pool of items). When using the total scores in a longitudinal analysis, the researcher has to make a choice between the IRT scale scores and the thetas. If one wishes to make interpretations about the amount of gains in terms of the additional number of items passed, the scale score is the most appropriate. For an analysis that is primarily targeting children in the upper end of the ability distribution (or the lower end), thetas might provide more discrimination between individual children. For most analyses, results based on scale scores or thetas will be very similar.

Table 8-1 presents the fall- and spring-kindergarten reading means, standard deviations and correlations, in the scale score metric, for the subsample used in this analysis. Inspection of table 8-1 indicates the standard deviations increase from fall- to spring-kindergarten. The increase in standard deviations suggests the potential of observing a "fan spread effect" (Campbell & Erlebacher 1970). The correlation of .03 between initial status and gain suggests little or no linear relationship between initial status and amount of gain. It appears that the adaptive test worked as expected, minimizing floor and ceiling effects. This low correlation suggests that the standard analysis of covariance approach that controls for initial status, and the repeated measures approach that analyzes the simple difference scores, yield very similar results.

Table 8-1.—Means, standard deviations, and correlations of reading scale scores fall- and spring-kindergarten

Category	Scale score round 1	Scale score round 2	Gain
Mean	22.65	32.56	9.91
Standard deviation	8.55	10.81	6.32
Scale score round 1	1.00		
Scale score round 2	.81	1.00	
Gain	.03	.61	1.00

Table 5-8 in chapter 5 shows nearly identical average gains from fall- to spring-kindergarten (round 1 to round 2) of about 10 scale score points for all subgroups, even though the *mean* scores for the subgroups are quite different. The same pattern appears for the theta scores in table 5-5: subgroups have very different means but similar average gains of about .80 to .90 in the theta metric. Individual or subgroup differences in the *amount* of gain given a relatively standard treatment (the year of kindergarten schooling) can be relatively trivial compared to differences in the average scores, that is, *where* on the developmental scale the gains are taking place. Thus analysis of total scale score gain tells only part of the story.

8.2 Proficiency Probabilities to Measure Longitudinal Change

The measurement approach used in developing the ECLS–K direct cognitive measures was to develop an IRT (Lord 1980) vertically equated scale using an adaptive test with multiple criterion referenced points along that vertical scale. The criterion referenced points model critical stages in the development of reading skills. In addition criterion referenced points serve two purposes at the individual level: (1) They provide information about changes in level of the child's mastery or proficiency, and (2) they provide information about where on the scale the child's gain is taking place. This latter piece of information about the child will be referred to as the location of maximum gain.

This chapter shows how one can identify the location of maximum gain on a hierarchical scale that is criterion referenced to represent five critical steps in the development of early reading skills. (Although not carried out here, the same procedure can be applied to the multiple criterion referenced mathematics scale.) Along with classifying children based on how much they are growing in relation to each of the five criterion referenced points on the growth curve, one can attempt to predict from background variables whether a child is making his/her gains at a selected critical point on the developmental scale. The scores used in this analysis are the proficiency probability scores described in chapter 3 and in the ECLS–K users manuals.

8.3 Method

The first step in the analysis was to use the criterion referenced points and the IRT model (Lord 1980) to locate where on the IRT ability (theta) scale each child was making his or her largest gain.

Figure 8-1 shows the location on the developmental scale of the five clusters of items marking the critical points on the scale. The numbers on the scale correspond to the ability level at which the probability of mastery of the particular skills reached 50 percent. Given a child's latent trait measure, theta, one could estimate the probability that a given child had mastered the knowledges associated with each of the critical points on the growth curve. These probabilities are the proficiency-level probability scores available on the public use data file.

Figure 8-1.—Proficiency levels theta scale

Letter recognition words in context (level 1)	Beginning sounds (level 2)	Ending sounds (level 3)	Sight comprehension of words (level 4)	Comprehension of words in context (level 5)
|	|	|	|	|
-.89	-.02	.45	1.34	1.85

The critical concept of location of maximum gain, which identifies at which of the five critical stages in development the child is making his or her maximum gain, is estimated in the following way. Differences between round 1 and round 2 in the probability of mastery are computed for each of the five proficiency levels. The largest difference marks the mastery level where the largest gain for a given child is taking place. This is the location of maximum gain for that child. For example, if the largest difference in probabilities of mastery for Sheila occurs at proficiency level 3, one can say that Sheila is making her largest gains in the mastery of ending sounds. This simple algorithm is used to find a unique location of maximum gain for each child. Having identified five mutually exclusive groups of children according to the proximity of their gains to each of the five critical points on the scale, one can treat the different types of gains as qualitatively different outcome measures to be explained by background and process variables.

Multilevel logistic regressions (Bryk & Raudenbush 1992, Snijders & Bosker 1999) were run to describe differences in profiles of those children who were gaining in level 4 and level 5 skills, contrasted with children who were making their maximum gains in the levels marking the middle and lower end of the developmental scale. Levels 4 and 5 were chosen because gains in this area of the scale have to do with beginning reading, while levels 1 to 3 are primarily measuring prereading mechanics. In the multilevel logistic regressions the dependent variable was coded "1" if the child was making his/her maximum gains at level 4 or 5, and "0" if the maximum gain was at level 1 to 3. The multilevel logistic regressions were estimated using quasi-likelihood estimators available in the MLWIN software (Goldstein et al. 1998). The binary dependent variable was analyzed with school at level 2 and child within school at level 1. All explanatory variables were fixed, and only the intercepts were considered random. These multilevel logistic regressions speak directly to the question of whether children who were changing at or above this critical point on the developmental scale come from different backgrounds and attended different types of schools than those children changing below this point (i.e., growing in their prereading mechanical skills).

In addition multilevel regressions were run on scale score gains that explicitly take into consideration where the gains were taking place on the scale. These results were then compared to the traditional approaches to measuring change.

8.4 Results

8.4.1 Gender and Location of Maximum Gain

Figure 8-2 shows a plot of fall- to spring-kindergarten reading gains by gender in the total scale score metric adjusted for age at first testing, time lapse between testing, and parents' education. Inspection of the plot in figure 8-2 indicates that girls were more advanced at entry to fall-kindergarten and increased their advantage on retesting. In terms of the classical repeated measure analysis, there was a significant gender by time of testing interaction indicating differential gain (F=38.07; p=.00). Similarly, the classical ANCOVA with the pretest, parents' education, and the two age-related variables as covariates yields adjusted spring reading means that significantly favor girls (F=38.14; p=.00). As expected, the ANCOVA and repeated measures yielded almost identical "F" tests since the gain scores were essentially uncorrelated with initial status.

Figure 8-3 presents a clustered histogram showing the location of maximum gains by gender. Inspection of figure 8-3 indicates that slightly over one-third of the boys were making their maximum gains in Letter Recognition, while about one-quarter of the girls did the same. Furthermore, girls were more likely than boys to be making their maximum gains in the higher-level proficiencies such as Ending Sounds. Only a very small percentage of both boys and girls made their greatest gains at the highest level (Comprehension of Words in Context) during the kindergarten year.

Figure 8-2.—Fall to spring reading gains by gender, adjusted for age and parent education

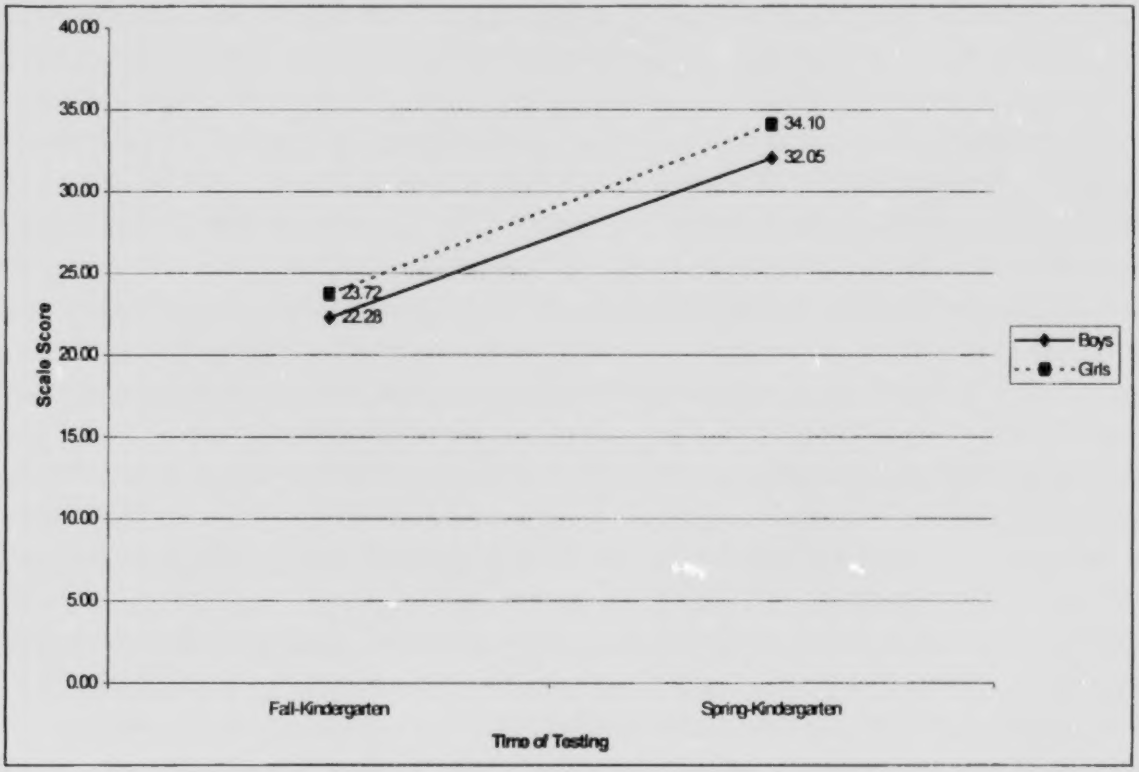
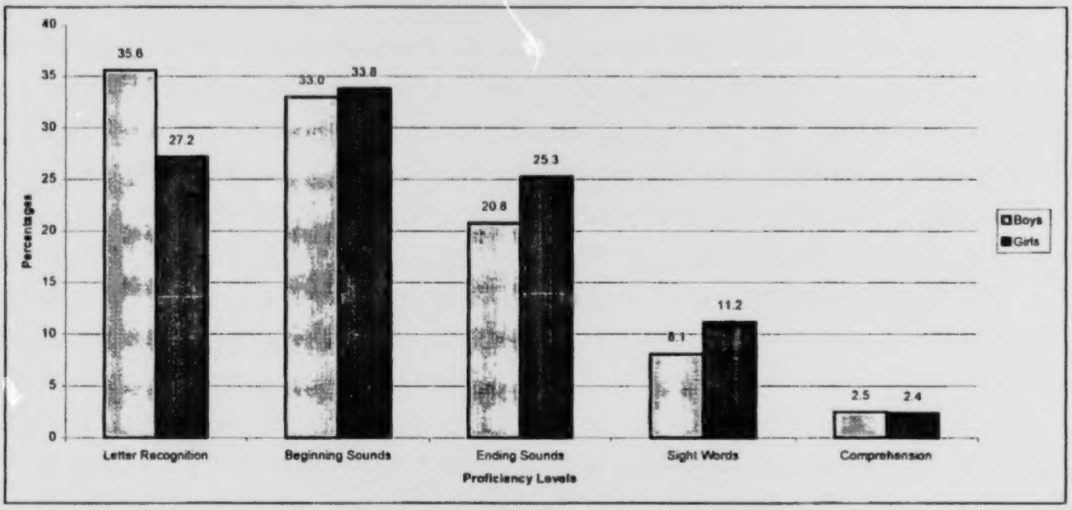


Figure 8-3.—Location of maximum gain by gender, fall- to spring-kindergarten



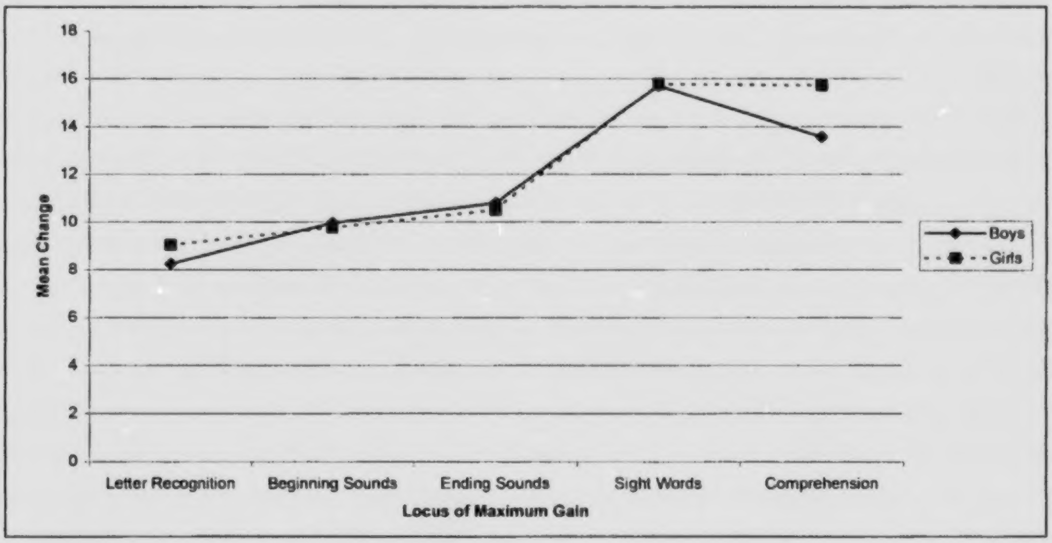
The next step in the analysis was to run the multilevel logistic regression described earlier with those children that were making their maximum gains in beginning reading (levels 4 to 5) coded "1", while those children making their maximum gains in prereading mechanics were coded "0". In the first set of these multilevel regressions, gender was the explanatory variable of specific interest, and parents' education, age at first testing, and time lapse between testing were used as covariates. Table 8-2 presents the gender logistic partial regression weights and their associated odds ratios and tests of significance. The last column of table 8-2 presents the reduction in the between-school variance that was due to (1) maturation as measured by age at first testing and the time lapse to the second testing, (2) the block of dummy variables contrasting various parents' educational levels with the base level (less than high school), and (3) the variable of interest, gender. Inspection of column 3, the odds ratios, indicates that girls were almost 1½ times (1.42) more likely than boys to be making their maximum gains at the two highest levels of the reading developmental scale. The block of parents' education contrasts reduced the between-school variance about 32 percent from that present in the null model (i.e., the intercept-only model). Clearly there were large differences between schools with respect to parents' education and those differences were related to where on the scale the children were making their gains. The odds ratios associated with those children whose parents have postgraduate schooling shows a disproportionately larger increase compared to the other contrasts, suggesting the possibility of a nonlinear relationship between years of parents' schooling and where children make their maximum gains.

Table 8-2.—Gender multilevel binomial analysis of gains at levels 4 to 5 versus gains at levels 1 to 3

Explanatory variables (1)	Logistic regression weights (2)	Odds ratios for gender equation (3)	"t" Statistics (P-value) (4)	Reduction in between-school variance (5)
Age fall-kindergarten	.07	1.07	11.5 (.00)	0.00%
Age-change	.04	1.04	.69 (.50)	
High school graduate	1.27	3.56	5.90 (.00)	
Some college	1.70	5.46	7.93 (.00)	
College graduate	2.47	11.80	11.47 (.00)	32.16%
Postgraduate studies	2.92	18.46	13.37 (.00)	
Girls	.35	1.42	6.62 (.00)	0.00%

These analyses were primarily concerned with where on the developmental scale the children are making their maximum gains. Figure 8-4 contrasts the genders on the *amount* of their gains by groups defined by their location of maximum gains. The plots in figure 8-4 represent the adjusted cell means from a two-way ANCOVA with gender and location of maximum gain as the design factors and scale score gains as the dependent variable. The covariates were age at first testing, time lapse between testing, and parents' education. The two-way interaction of gender by location of maximum gain was statistically significant (F=9.00; p=.00). Inspection of figure 8-4 indicates that while girls and boys were *equally represented* at the highest level (level 5), girls were making *significantly greater gains* at this level. Conversely, girls were underrepresented at level 1, but those girls who were making their maximum gains at this level were making greater gains on the total scale score metric than are boys who were at this level of development.

Figure 8-4.—Adjusted mean changes on the total score scale by gender and proficiency level



8.4.2 School Sector and Location of Maximum Gain

Figure 8-5 presents the fall to spring reading gains by school sector adjusted for age, time lapse between testing, and parents' education. Inspection of figure 8-5 suggests that in terms of the total scale score metric, schools from the different sectors had quite different initial status with respect to their children's developmental level, and they maintained the same relative positions on retesting in the spring. In terms of the traditional repeated measures approach, the school sector by trial interaction (F=. 50; p=. 60) was not significant. An ANCOVA analysis with the fall-kindergarten test scores, age, time lapse, and parents' education as covariates yielded almost exactly the same results (F=. 53; p=. 58). That is, there was no difference in the amount of gain for children from the different school sectors in terms of the total scale score metric.

Figure 8-6 presents a histogram detailing where on the scale the children from different school sectors were making their maximum gains. It indicates that there were quite different patterns with respect to where the gains were taking place. Inspection of figure 8-6 indicates that a full one-third of the public school children were making their maximum gains at the lowest level skill (Letter Recognition), while only 10 percent of the private non-Catholic school children were making their maximum gains at this level. Conversely, one-third of the private non-Catholic school children compared with about 11 percent of the public school children were making their maximum gains at the early reading skills, Sight Words and Comprehension of Words in Context (levels 4 and 5 respectively). Contrasts between the Catholic and public school children with respect to the location of maximum gain indicates that two-thirds of the Catholic school children were making their maximum gains in Beginning Sounds and Ending Sounds, while a similar proportion of the public school children were making their maximum gains in Letter Recognition and Beginning Sounds.

Figure 8-5.—Fall- to spring-kindergarten reading gains by school sector, adjusted for age and parent education

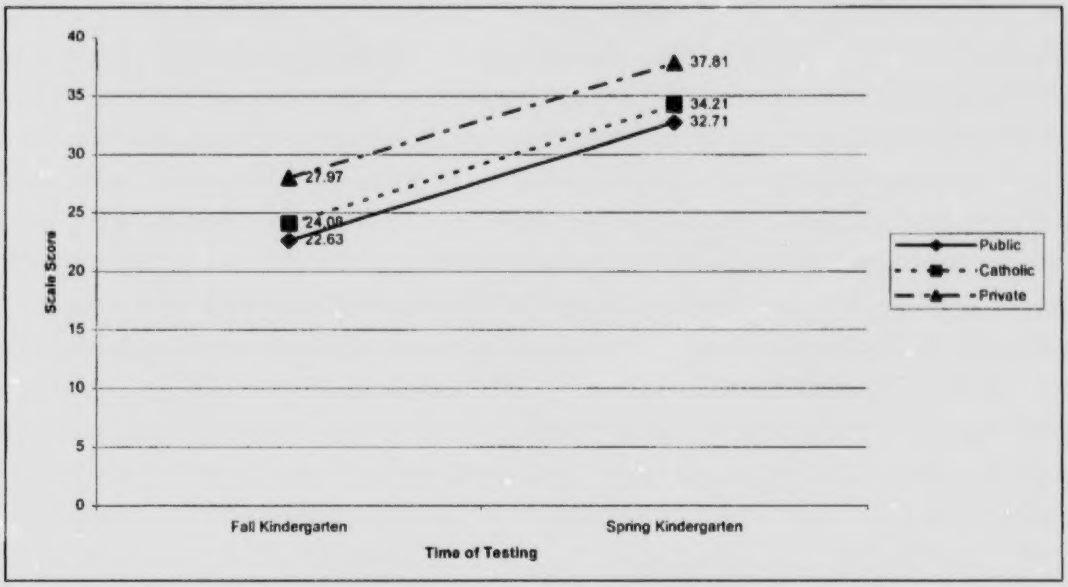


Figure 8-6.—Location of maximum reading gain by school type, fall- to spring-kindergarten

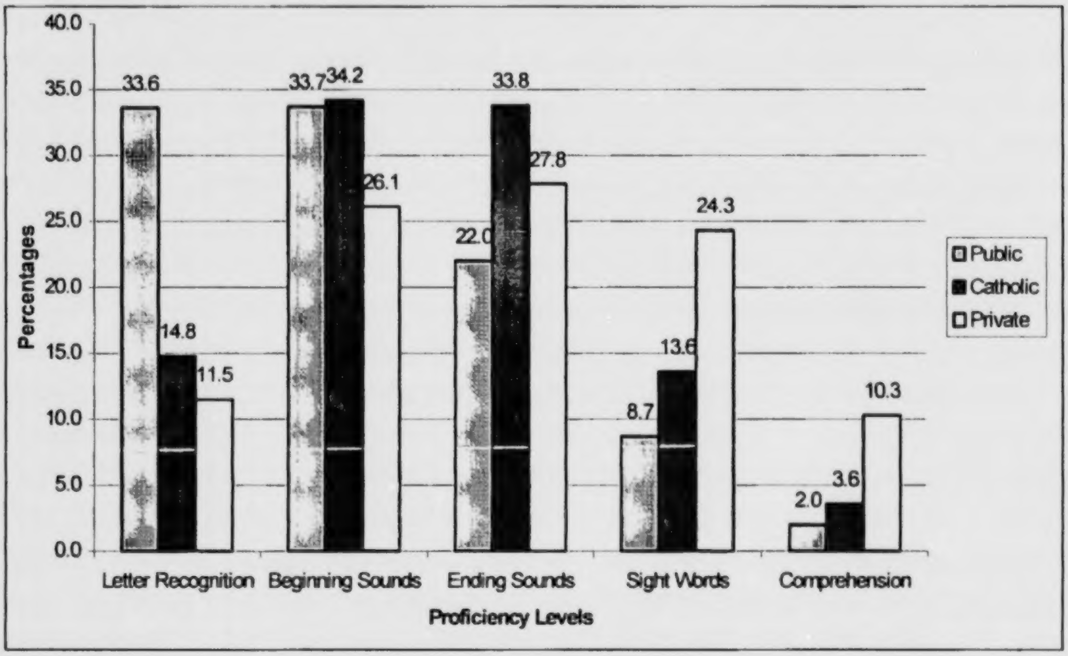


Table 8-3 presents the multilevel logistic partial regression weights relating the school sector explanatory variables controlling for age at time of first testing, time lapse between testing, and parents' education to the dichotomous outcome of whether the child was making his/her maximum gains at levels 4 to 5 versus the lower three levels dealing with prereading mechanics. In this multilevel logistic regression analysis, the public schools were the base or contrast group. This school sector analysis parallels the gender analysis described earlier. Inspection of the odd ratios in column 3 of table 8-3 indicates that the private non-Catholic school children were almost three times as likely as public school children to be making their maximum gains in the higher level beginning reading skills (levels 4 to 5). There was no significant difference between the Catholic school children and the public school children with respect to this dichotomous criterion.

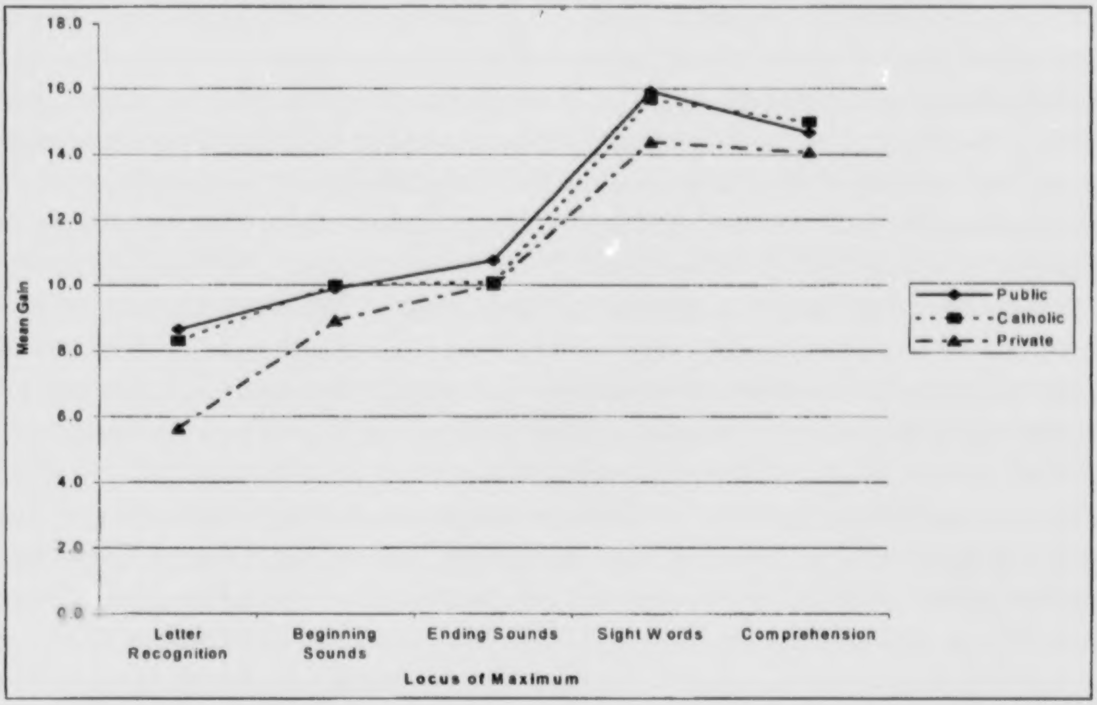
Table 8-3.—School sector binomial multilevel analysis of gains at levels 4 to 5 versus gains at levels 1 to 3

Explanatory variables (1)	Logistic regression weights (2)	Odds ratios for gender equation (3)	"z" Statistics (P-value) (4)	Reduction in between-school variance (5)
Age fall-kindergarten	.07	1.07	11.33 (.00)	0.00%
Age-change	.05	1.05	.96 (.32)	
High school graduate	1.27	3.56	5.79 (.00)	32.16%
Some college	1.68	5.38	7.68 (.00)	
College graduate	2.43	11.31	11.08 (.00)	
Postgraduate studies	2.84	17.10	12.73 (.00)	
Catholic	.14	1.15	1.11 (.26)	
Private non-Catholic	1.07	2.91	7.02 (.00)	6.62%

A comparison of the reductions in the between-school variance (column 5) suggests that the introduction of the school sector variables did reduce the between school variance 6 percent. The fact remains that the major part of the explainable between-school variance in cognitive growth in beginning reading skill (levels 4 and 5) was associated with parents' education and even more specifically, college educated and postgraduate parents. This result is consistent with the notion that children from homes with college educated parents are more likely to enter kindergarten already knowing their prereading skills and, thus, are making their gains in level 4 and 5 skills.

Figure 8-7 shows the amount of gain in total scale score points taking place for groups defined by their location of maximum gain crossed with school sector. This figure shows graphically the results of a school sector by location of maximum gain ANCOVA. Table 8-4 shows the parallel results from a multilevel analysis with gains in the total scale score as the dependent variable and dummy variables for school sector and location of maximum gain.

Figure 8-7.—Adjusted mean gains on the total score scale by school sector and location of maximum gain



Column 2 of table 8-4 gives the partial regression weights associated with each explanatory variable when all variables were entered (i.e., full model). The overall fit statistic is presented in column 4 for the age block, parents' education block, location of maximum gain block, and finally the school sector block. It is the reduction in the overall fit statistic as each succeeding block is added to the model that is of interest. The fifth column shows the cumulative reduction in the between-school variance as each block is added to the model. The relative stability of the intra-class correlation in the presence of relatively large reductions in the between-school variance, as in the case of the location of maximum gain block, suggests that proportionately equivalent reductions were made in the between-individual variance. Inspection of column 4 indicates that the block associated with parental education contributes little to the explanation of the amount of gain. As shown earlier (tables 8-2 and 8-3), parents' education explains where the gains are being made, but not the amount of gain. That is, parents' education has much to do with the child's status at entry to kindergarten, and thus indirectly on *where* he or she is making his/her maximum gain. But since status and gain have a zero correlation, parents' education will also have little or no correlation with *amount* of gain.

Table 8-4.—Multilevel analysis of raw gains by school sector and location of maximum gain (locmg)

Explanatory variables	Regression coefficients (SE) full model	Probability	-2ln Likelihood	Reduction in between-school variance (cum.)	Intra-class correlation
Age first testing	.06 (.01)	.00	88432	17.45%	.12
Time lapse	1.10 (.10)	.00			
High school graduate	.35 (.19)	.06	88349	.63%	.12
Some college	.57 (.20)	.00			
College graduate	.15 (.22)	.50			
Postgraduate studies	.15 (.25)	.54			
Locmg2	1.19 (.12)	.00	84011	21.6%	.11
Locmg3	1.82 (.14)	.00			
Locmg4	6.84 (.19)	.00			
Locmg5	5.80 (.32)	.00			
Catholic	-.37 (.29)	.20	83998	.81%	.11
Private non-Catholic	-1.18 (.38)	.00			

Inspection of figure 8-7 suggests that when location of maximum gain is controlled, public school and Catholic school children were consistently gaining as much or more than the children in the private non-Catholic schools. In fact, while the interaction was not significant (chi-square =10; df=6, p=.12), the school sector main effect was significant (chi-square=13; df=2,p=.00) with both the public and Catholic children showing significantly greater gains than the private non-Catholic school children. This appears to contradict the results discussed earlier and shown in figure 8-5. This is a result of the private non-Catholic school children being disproportionately over-represented in the groups making their maximum gains at levels 4 and 5 (see figure 8-6), where the average *amount* of gain was greater, while more of the public school children were making their maximum gains at levels 1 to 3. That is, less than 10 percent of the public school children and 17 percent of the Catholic school children made their maximum gains at levels 4 and 5. However, those public and Catholic school children that did make their maximum gains at level 4 and 5 outperformed their counterparts from the private non-Catholic schools.

When one controls for where the gains take place (i.e., the location of maximum gain), there is a significant reduction in the between-school variance (21.6 percent as shown in table 8-4), and one gets different results. While the interaction between school sector and location of maximum gain was not significant, there was some evidence that children entering kindergarten with only level 1 or lower skills may gain more at the public or Catholic schools. The results graphed in figure 8-7, which are consistent with the significant main effects shown in table 8-4, suggest that public and Catholic school kindergartens have a positive influence on reading gains at all levels of the developmental scale.

8.5 Alternative Measure of Overall Gain

As pointed out earlier, the complex patterns that can occur with respect to gain scores are not always properly summarized in a single overall measure of gain, whether they are raw gains from repeated measures or residualized gains from ANCOVA with the pretest as a covariate. The use of adaptive tests makes this even more complicated. It is advantageous to have an additional summary gain score that can take into consideration the amount of gain, as well as where on the scale the gain is occurring. The percent of maximum possible gain is suggested as an alternative single summary scoring procedure that takes into consideration where the gain is taking place on the developmental scale. This scoring system implicitly assumes gains at the upper end of the scale are more important than at the lower levels. The logic for this is as follows. At a given point in time in a developmental process, such as learning to read, those children who are making their gains at the upper end of the developmental scale will be better positioned for further advancement in reading skills. In addition, as they become more skilled in reading comprehension, they will be able to use reading as a tool in mastering other school related skills. Equation (8.1) estimates the percent of maximum possible gain as follows:

$$\hat{Y}_{gi} = \left[(y_{i2} - y_{i1})/(y_{\max} - y_{i1})\right]x100. \tag{8.1}$$

where $\hat{Y}_{gi}$ = percent of maximum gain for individual i
y_{i2} = total scale score at time 2 for individual i
y_{i1} = total scale score at time 1 for individual i
$y_{\max}$ = maximum possible total scale score on the item pool.

The percentage of maximum possible gain as defined in (8.1) also has the potential for helping to minimize the impact of ceiling effects if they should occur. Percentage of maximum gain can be viewed as a gain score variation on the POMP score suggested by Cohen et al. (1999) for measuring status at a single point in time.

Table 8-5 compares multilevel results for raw gains with the percentage of maximum gains. Inspection of table 8-5 suggests that the block of parents' education variables becomes considerably more important when the outcome is percentage of maximum gain. This is not surprising since percentage of maximum gain includes a component related to where on the scale the gain takes place. More importantly, the signs of the school sector variables have gone from negative to positive and the private non-Catholic regression weight comes close to significance. These findings come closer to what is known about the performance of children in the private non-Catholic schools. That is, they are disproportionately represented in the group making their maximum gains at the upper end of the scale. When the next followup, spring-first grade, is analyzed this measure of gain may be more effective in summarizing what is really happening. The present longitudinal scale is fixed through the spring of first grade and thus more children will be approaching the upper end of the scale at that time.

8.6 Conclusions

The methodology used in this analysis used adaptive tests with multiple criterion-referenced points that mark critical points in the early reading developmental process. Emphasis was placed on where on the vertical scale gain was taking place, as well as the amount of gain. The results show the following:

- Traditional approaches to measuring gain found no differences between school sectors. However, when the location of maximum gain was explicitly controlled, children at both public and Catholic schools showed significantly greater gains than did their other private school counterparts. While parents' education was highly related to where the gains were being made, and thus, of course, directly to reading skills at kindergarten entry, controlling for pre-test scores and parents' education may be misleading. The reasons for this were tied up in the distributional differences with respect to where the gains were taking place, as well as a non-linear relationship between *amount* of gain and *where* on the scale that gain was taking place. Children in public and Catholic school kindergartens made gains at all levels of the reading proficiency scale. On average the private non-Catholic school children entered kindergarten with more advanced reading skills than their counterparts in the other school sectors. Thus they were over-represented with respect to growth at the upper end of the scale associated with beginning reading, and because of their advanced skills may have the potential of widening the gap in the future.

- Girls began kindergarten at a younger age and with better prereading skills than did boys.

- On the whole, girls gained more than boys in the total scale score metric, and this finding was independent of the analytic method used.

- Boys and girls differed on where on the scale they made their gains. Boys were almost twice as likely as girls to be making their gains in the lowest level prereading skill, Letter Recognition.

- Girls were more likely than boys to be making their gains in the areas of the scale related to Ending Sounds (level 3) and Sight Words (level 4).

- Girls and boys have about equal representation among children who were gaining at the upper end of the scale defined by level 5 skills (Comprehension of Words in Context).

- Parents' education was closely related to where the gains were taking place on the developmental scale but had little relation to the amount of gain once the location of maximum gain was entered into the model.

- On average, children in public schools had the lowest reading skills on entry to kindergarten, followed by children entering Catholic schools, with children entering private non-Catholic schools having the highest reading skills.

- Children attending public schools were much more likely than children attending Catholic or private non-Catholic schools to be gaining on level 1 tasks (Letter Recognition) during the kindergarten year.

- Children attending Catholic schools were much more likely to be making their gains in level 3 skills (Ending Sounds) than children attending either public or private non-Catholic schools.

- Children attending private non-Catholic schools were much more likely than children at public or Catholic schools to be making their gains in level 4 (Sight Words) and level 5 (Comprehension of Words in Context). This differential in favor of private non-Catholic schools was particularly large for the level 5 tasks.

REFERENCES

American Association for the Advancement of Science. (1995). *Benchmarks for science literacy.* [on-line]. Available: www.project2061.org.

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical linear models, applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Campbell, D.T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 15, 546-53.

Campbell, D.T., and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Campbell, D.T., and Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.) *Compensatory education: A national debate.* New York: Brunner/Hazel.

Cohen, P., Cohen, J., Aiken, L., and West, S.G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research,* 34, 3, 315-346.

Cole, N.S., and Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.) *Educational Measurement,* (3rd Ed., pp. 201-219). New York: American Council on Education/Macmillan.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Duncan, S.E., and De Avila, E. (1998). *PreLAS 2000.* Monterey, CA: CTB/McGraw-Hill.

Duncan, S.E., and De Avila, E. (1986). *PreLAS.* Monterey, CA: CTB/McGraw-Hill.

Dunn, L.M., and Dunn, L.M. (1981). *Peabody Picture Vocabulary Test-Revised (PPVT-R).* Circle Pines, MN: American Guidance Services, Inc.

Ginsburg, H.P., and Baroody, A.J. (1990). *The Test of Early Mathematics (TEMA-2).* Austin, TX: PRO-ED, Inc.

Goldstein, H. (1995). *Multilevel statistical models. 2nd Ed.* London: Edward Arnold.

Goldstein, H., Rashback, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., and Healy, M. (1998). *A user's guide to MLwinN.* London: Multilevel Models Project, Institute of Education, University of London.

Gresham, F., and Elliot, S. (1990). *Social skills rating system.* Circle Pines, MN: American Guidance Services, Inc.

Guttman, L.A. (1954). A new approach to factor analysis: The radix. In, P.F. Lazersfeld (Ed.): *Mathematical thinking in the social sciences.* New York: Columbia University Press.

Harcourt Brace. (1995). *Science anytime.* Orlando, FL: Author.

Holland, P.W., and Thayer, D.T. (1986). *Differential item function and the Mantel-Haenszel procedure.* (ETS Research Report No. 86-31). Princeton, NJ.

Huttenlocher, J., and Levine, S.C. (1990). *The Primary Test of Cognitive Skills (PTCS).* New York: CTB/McGraw Hill.

Ingels, S., et al. (1997). *Field test report: National Longitudinal Study of 1988 (Base Year).* Chicago, IL: NORC, University of Chicago.

Kaufman, A.S., and Kaufman, N.L. (1985). *Kaufman Test of Educational Achievement (KTEA).* Circle Pines, MN: American Guidance Services, Inc.

Linacre, J.M., and Wright, B.D. (2000). *A user's guide to Winsteps Ministep Rasch model computer programs.* Chicago, IL: MESA Press.

Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Publishers.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute,* 22, 719-748.

Markwardt, F.C., Jr. (1989). *Peabody Individual Achievement Test-Revised (PIAT-R).* Circle Pines, MN: American Guidance Services, Inc.

Meisels, S.J., and Perry, N.E. (1996). *How accurate are teacher judgments of student's academic performance?* (Working Paper No. 96-08). Washington, DC: National Center for Education Statistics.

Meisels, S.J., Marsden, D.B., Wiske, M.S., and Henderson, L.W. (1997*). The Early Screening Inventory – Revised.* Ann Arbor, MI: Rebus, Inc.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359-381.

Mislevy, R.J., and Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models.* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R.J., et al. (1992). Scaling procedures in NAEP. *Journal of Education Statistics*, 17, 131-154.

Montgomery, D. (1997). *Identification of an English proficiency measure for the Early Childhood Longitudinal Study*. Palo Alto, CA: American Institutes for Research.

Muraki E.J., and Bock, R.D. (1987). *BIMAIN: A program for item pool maintenance in the presence of item parameter drift and item bias.* Mooresville, IN: Scientific Software.

Muraki E.J., and Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data [computer program].* Chicago, IL: Scientific Software, Inc.

National Academy of Sciences. (1995). *National science education standards.* Washington, DC: Author.

National Assessment Governing Board (NAGB). (1994a). *Reading Framework for the 1992 and 1994 National Assessment of Educational Progress.* Washington, DC: Government Printing Office.

National Assessment Governing Board (NAGB). (1994b). *Geography Framework for the 1994 National Assessment of Educational Progress.* Washington, DC: Government Printing Office.

National Assessment Governing Board (NAGB). (1996a). *Mathematics Framework for the 1996. National Assessment of Educational Progress.* Washington, DC: Government Printing Office.

National Assessment Governing Board (NAGB). (1996b). *Science Framework for the 1996 National Assessment of Educational Progress.* Washington, DC: Government Printing Office.

National Center for Education Statistics. (2001). *ECLS-K First Grade Restricted-Use User's Manual* (NCES 2001–101).

National Center for Education Statistics. (August, 2000). *ECLS-K Restricted-Use Base Year User's Manual.* (NCES 2001–097).

National Center for Education Statistics. (February, 2000). *ECLS-K Base Year Public-Use User's Manual.* (NCES 2001–029).

National Center for Education Statistics. (November, 2001). *ECLS-K First Grade Restricted-Use Electronic Codebook* (NCES 2002–128).

National Council for the Social Studies. (1994). *Curriculum Standards for Social Studies: Expectations of Excellence*. Washington, DC.

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: Author.

National Research Council (1996). National Science Education Standards. Washington, DC: National Academy Press.

Ramsey, W.L. (1986). *Holt science.* New York: Holt, Rinehart and Wilson.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute.

Reid, D.K., Hresko, W.P., and Hammill, D.D. (1981). *Test of Early Reading Ability (TERA-2).* Austin, TX: PRO-ED, Inc.

Rock, D.A., and Pollack, J. (1987). The Cognitive test battery. In S.J. Ingels, et al., *Field test report: National Education Longitudinal Study of 1988 (Base Year).* Chicago, IL: NORC, University of Chicago.

Rock, D.A., et. al. (1985). *Psychometric analysis of the NLS-72 and the High School And Beyond test batteries*. (NCES Report No.85-217). Washington, DC: National Center for Education Statistics.

Rock, D.A., et. al. (1995). *Psychometric report for the NELS:88 base year test battery*. (NCES Report No.95-382). Washington, DC: National Center for Education Statistics.

Scott-Foresman. (1994). *Discover the wonder*. Glenview, IL: Author.

Silver Burdett & Ginn. (1991). *Science Horizons*. Lexington, MA: Author.

Smith, R.M., Schumacker, R.E., and Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2(1), 66-78.

Snijders, T., and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications.

STATA Corporation. (2000). *STATA Statistical Software: Release 7*. College Station, TX: STATA Corporation.

Woodcock, R.W., and Bonner, M. (1989). *The Woodcock-Johnson Tests of Achievement-Revised*. Itasca, IL: Riverside Publishing Company.

Wright, B.D. (1999). Fundamental measurement for psychology. In S. Embretson and S. L. Hershberger (Eds.) *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Wright, B.D., and Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Yamamoto, K., and Mazzeo, J. (1992). Item Response Theory: Scale linking in NAEP. *Journal of Education Statistics*, 17, 155-173.

Appendix A

Summary of 1989 NCTM Mathematics Curriculum Standards for Grades Kindergarten to Fourth

APPENDIX A. SUMMARY OF 1989 NCTM MATHEMATICS CURRICULUM STANDARDS FOR GRADES KINDERGARTEN TO FOURTH

Standard 1. Mathematics as Problem Solving

In grades kindergarten to fourth, the study of mathematics should emphasize problem solving so that students can:

- Use problem-solving approaches to investigate and understand mathematical content;
- Formulate problems from everyday and mathematical situations;
- Develop and apply strategies to solve a wide variety of problems;
- Verify and interpret results with respect to the original problem; and
- Acquire confidence in using mathematics meaningfully.

Standard 2. Mathematics as Communication

In grades kindergarten to fourth, the study of mathematics should include numerous opportunities for communication so that students can:

- Relate physical materials, pictures, and diagrams to mathematical ideas;
- Reflect on and clarify their thinking about mathematical ideas and situations;
- Relate their everyday language to mathematical language and symbols; and
- Realize that representing, discussing, reading, writing, and listening to mathematics are a vital part of learning and using mathematics.

Standard 3. Mathematics as Reasoning

In grades kindergarten to fourth, the study of mathematics should emphasize reasoning so that students can:

- Draw logical conclusions about mathematics;
- Use models, known facts, properties, and relationships to explain their thinking;
- Justify their answers and solution processes;
- Use patterns and relationships to analyze mathematical situations; and
- Believe that mathematics makes sense.

Standard 4. Mathematical Connections

In grades kindergarten to fourth, the study of mathematics should include opportunities to make connections so that students can:

- Link conceptual and procedural knowledge;
- Relate various representations of concepts or procedures to one another;
- Recognize relationships among different topics in mathematics;
- Use mathematics in other curriculum areas; and
- Use mathematics in their daily lives.

Standard 5: Estimation

In grades kindergarten to fourth, the curriculum should include estimation so that students can:

- Explore estimation strategies;
- Recognize when an estimate is appropriate;
- Determine the reasonableness of results; and
- Apply estimation in working with quantities, measurement, computation, and problem solving.

Standard 6. Number Sense and Numeration

In grades kindergarten to fourth, the mathematics curriculum should include whole number concepts and skills so that students can:

- Construct number meanings through real-world experiences and the use of physical materials;
- Understand our numeration system by relating counting, grouping, and place-value concepts;
- Develop number sense; and
- Interpret the multiple uses of numbers encountered in the real world.

Standard 7. Concepts of Whole Number Operations

In grades kindergarten to fourth, the mathematics curriculum should include concepts of addition, subtraction, multiplication, and division of whole numbers so that students can:

- Develop meaning for the operations by modeling and discussing a rich variety of problem situations;
- Relate the mathematical language and symbolism of operations to problem situations and informal language;
- Recognize that a wide variety of problem structure can be represented by a single operation; and
- Develop operation sense.

Standard 8. Whole Number Computation

In grades kindergarten to fourth, the mathematics curriculum should develop whole number computation so that students can:

- Model, explain, and develop reasonable proficiency with basic facts and algorithms;
- Use a variety of mental computation and estimation techniques;
- Use calculators in appropriate computational situations; and
- Select and use computation techniques appropriate to specific problems and determine whether the results are reasonable.

Standard 9. Geometry and Spatial Sense

In grades kindergarten to fourth, the mathematics curriculum should include two- and three-dimensional geometry so that students can:

- Describe, model, draw, and classify shapes;
- Investigate and predict the results of combining, subdividing, and changing shapes;
- Develop spatial sense;
- Relate geometric ideas to number and measurement ideas; and
- Recognize and appreciate geometry in their world.

Standard 10. Measurement

In grades kindergarten to fourth, the mathematics curriculum should include measurement so that students can:

- Understand the attributes of length, capacity, weight, area, volume, time, temperature, and angle;
- Develop the process of measurement;
- Make and use estimates of measurement; and
- Make and use measurements in problem and everyday situations.

Standard 11. Statistics and Probability

In grades kindergarten to fourth, the mathematics curriculum should include experiences with data analysis and probability so that students can:

- Collect, organize and describe data;
- Construct, read, and interpret displays of data;
- Formulate and solve problems that involve collecting and analyzing data; and
- Explore concepts of chance.

Standard 12. Fractions and Decimals

In grades kindergarten to fourth, the mathematics curriculum should include fractions and decimals so that students can:

- Develop concepts of fractions, mixed numbers, and decimals;
- Develop number sense for fractions and decimals;
- Use models to relate fractions to decimals and to find equivalent fractions;
- Use models to explore operations on fractions and decimals; and
- Apply fractions and decimals to problem situations.

Standard 13. Patterns and Relationships

In grades kindergarten to fourth, the mathematics curriculum should include the study of patterns and relationships so that students can:

- Recognize, describe, extend, and create a wide variety of patterns;
- Represent and describe mathematical relationships; and
- Explore the use of variables and open sentences to express relationships.

Appendix B

Reading Assessment Detailed Content Classifications Used for Item Development

APPENDIX B. READING ASSESSMENT DETAILED CONTENT CLASSIFICATIONS USED FOR ITEM DEVELOPMENT

1. **Basic Skills**
    - Print Familiarity
    - Letter Recognition
    - Beginning Sounds
    - Ending Sounds
    - Short Vowels
    - Long Vowels
    - Rhyming Words
2. **Vocabulary**
    - Picture-Spoken Word Matching
    - Word Recognition
3. **Initial Understanding***
4. **Developing Interpretation***

---

* Each of the four types of comprehension skills is measured by three kinds of test items:

(1) Listening comprehension questions based on passages,

(2) Reading comprehension questions based on sentences, and

(3) Reading comprehension questions based on passages.

Listening comprehension items are included only in the kindergarten and first grade tests.

Reading comprehension items begin in first grade and continue through the fifth grade.

BEST COPY AVAILABLE

5. **Personal Reflection**[*]

6. **Critical Stance**[*]

---

[*] Each of the four types of comprehension skills is measured by three kinds of test items:

(1) Listening comprehension questions based on passages,

(2) Reading comprehension questions based on sentences, and

(3) Reading comprehension questions based on passages.

Listening comprehension items are included only in the kindergarten and first grade tests.

Reading comprehension items begin in first grade and continue through the fifth grade.

Appendix C

Summary of the 1996 National Research Council Grades Kindergarten to Fourth Content Standards in Science

APPENDIX C. SUMMARY OF THE 1996 NATIONAL RESEARCH COUNCIL GRADES KINDERGARTEN TO FOURTH CONTENT STANDARDS IN SCIENCE

The following outline is excerpted from *National Science Education Standards* published by the National Research Council in 1996. Detailed descriptions of the concepts and skills, along with numerous examples of lessons, are included in that volume.

Content Standard A: Science as Inquiry

As a result of activities in grades kindergarten to fourth, all students should develop:

- Abilities necessary to do scientific inquiry and
- Understanding about scientific inquiry.

Content Standard B: Physical Science

As a result of activities in grades kindergarten to fourth, all students should develop an understanding of:

- Properties of objects and materials;
- Position and motion of objects; and
- Light, heat, electricity, and magnetism.

Content Standard C: Life Science

As a result of activities in grades kindergarten to fourth, all students should develop an understanding of:

- The characteristics of organisms;
- Life cycles of organisms; and
- Organisms and environments.

Content Standard D: Earth and Space Science

As a result of their activities in grades kindergarten to fourth, all students should develop an understanding of:

- Properties of earth materials;
- Objects in the sky; and
- Changes in earth and sky.

Content Standard E: Science and Technology

As a result of activities in grades kindergarten to fourth, all students should develop:

- Abilities of technological design;
- Understanding about science and technology; and
- Abilities to distinguish between natural objects and objects made by humans.

Content Standard F: Science in Personal and Social Perspectives

As a result of activities in grades kindergarten to fourth, all students should develop understanding of:

- Personal health;
- Characteristics and changes in populations;
- Types of resources;
- Changes in environments; and
- Science and technology in local challenges.

Content Standard G: History and Nature of Science

As a result of activities in grades kindergarten to fourth, all students should develop understanding of:

- Science as a human endeavor.

Appendix D

ECLS Item Parameters and Item Fit by Rounds

Table D-1.—Reading item parameters and item fit by rounds

		IRT Parameters			Round 1				Round 2				Round *3				Round 4			
						P+				P+				P+				P+		
Reading	Test Form(s)	a	b	c	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
LETREC D	Routing	1.60	-0.63	0.00	17607	0.74	0.73	0.00	18938	0.94	0.94	0.00	5053	0.96	0.97	-0.01	16339	0.99	0.99	-0.01
LETREC F	Routing	1.82	-0.56	0.00	17606	0.72	0.72	0.00	18939	0.94	0.93	0.00	5053	0.96	0.97	-0.01	16339	0.99	0.99	0.00
LETREC T	Routing	1.71	-0.42	0.00	17594	0.66	0.66	-0.01	18937	0.92	0.91	0.01	5053	0.96	0.95	0.00	16339	0.99	0.99	0.00
LETREC M	Routing	1.61	-0.55	0.00	17604	0.71	0.70	0.00	18935	0.93	0.93	0.00	5053	0.95	0.96	-0.01	16338	0.99	0.99	0.00
BEG P	Routing	1.19	0.22	0.05	17607	0.47	0.46	0.01	18939	0.77	0.75	0.02	5053	0.84	0.84	0.00	16340	0.93	0.96	-0.02
BEG R	Routing	1.71	0.29	0.08	17600	0.42	0.44	-0.01	18936	0.79	0.77	0.02	5053	0.88	0.87	0.01	16340	0.96	0.97	-0.01
BEG L	Routing	1.74	0.36	0.08	17605	0.41	0.42	-0.01	18940	0.77	0.75	0.02	5053	0.87	0.85	0.02	16340	0.96	0.97	-0.01
BEG B	Routing	0.97	0.72	0.02	17609	0.32	0.31	0.01	18938	0.58	0.57	0.01	5052	0.68	0.68	0.00	16340	0.86	0.87	-0.01
END D	Routing	1.35	1.10	0.06	17602	0.22	0.22	0.00	18937	0.48	0.47	0.01	5052	0.61	0.60	0.01	16338	0.86	0.86	0.00
END P	Routing	1.30	0.93	0.07	17614	0.28	0.27	0.01	18939	0.55	0.54	0.01	5052	0.67	0.66	0.01	16338	0.88	0.89	-0.01
END L	Routing	1.82	0.68	0.09	17599	0.32	0.32	0.00	18930	0.66	0.65	0.01	5051	0.78	0.78	0.00	16338	0.94	0.95	-0.01
END F	Routing	1.38	0.70	0.07	17607	0.32	0.32	0.00	18938	0.64	0.62	0.02	5050	0.75	0.74	0.01	16338	0.91	0.93	-0.02
RUNS	Routing	2.43	1.43	0.00	11573	0.10	0.10	-0.01	17489	0.29	0.29	0.00	4823	0.44	0.45	-0.01	16186	0.87	0.84	0.03
DOWN	Routing	2.89	1.64	0.00	11581	0.06	0.06	0.00	17490	0.18	0.20	-0.02	4825	0.33	0.33	-0.01	16184	0.82	0.78	0.04
WENT	Routing	2.41	1.58	0.00	11574	0.08	0.08	0.00	17480	0.24	0.23	0.00	4820	0.36	0.37	-0.01	16185	0.81	0.79	0.02
JEEP	Routing	2.31	1.64	0.00	11576	0.08	0.07	0.01	17474	0.21	0.21	0.00	4827	0.32	0.34	-0.02	16184	0.77	0.76	0.01
BACKPACK	Routing	2.48	2.14	0.19	700	0.47	0.45	0.02	3353	0.47	0.45	0.02	1597	0.54	0.51	0.02	13463	0.68	0.68	0.00
RIDEBIKE	Routing	3.36	2.31	0.23	697	0.47	0.41	0.06	3283	0.42	0.39	0.02	1584	0.45	0.45	-0.01	13534	0.60	0.60	0.00
LISTEN	Routing	3.71	2.21	0.15	675	0.42	0.39	0.03	3090	0.39	0.38	0.01	1479	0.46	0.46	0.00	13191	0.63	0.63	0.00
SIZES	Routing	3.59	2.33	0.14	662	0.36	0.34	0.03	2980	0.35	0.33	0.03	1442	0.40	0.39	0.00	12576	0.56	0.56	-0.01
CEREAL	Low	0.63	-2.57	0.00	13350	0.91	0.91	0.00	6516	0.94	0.94	0.00	1062	0.95	0.95	0.00	617	0.91	0.95	-0.04
BEG BIKE	Low	1.59	-0.36	0.42	13347	0.73	0.73	0.00	6520	0.84	0.85	0.00	1062	0.87	0.87	0.00	618	0.88	0.90	-0.01
BEGIN	Low,Mid,High	0.57	-0.86	0.01	17611	0.69	0.68	0.01	18936	0.83	0.83	0.00	5051	0.87	0.87	0.00	16338	0.93	0.94	-0.01
NEXTLINE	Low,Mid,High	0.72	-0.16	0.00	17610	0.54	0.55	0.00	18939	0.76	0.76	0.00	5052	0.85	0.82	0.03	16339	0.92	0.92	-0.01
STORYEND	Low,Mid,High	0.82	-0.16	0.00	17615	0.57	0.55	0.02	18940	0.78	0.78	0.00	5049	0.84	0.84	0.00	16336	0.92	0.94	-0.02
CANDLE	Low	0.45	-3.70	0.23	13348	0.95	0.94	0.00	6519	0.95	0.96	-0.01	1061	0.96	0.96	0.00	617	0.94	0.97	-0.02
DECORATD	Low	0.45	-2.17	0.19	13307	0.84	0.84	0.00	6506	0.87	0.88	0.00	1060	0.87	0.89	-0.01	615	0.86	0.89	-0.04
POURINT	Low	0.50	-2.44	0.21	13327	0.89	0.88	0.01	6507	0.91	0.91	-0.01	1051	0.90	0.92	-0.03	612	0.87	0.93	-0.06
VEGETBLE	Low, Mid	0.47	-0.38	0.20	16942	0.66	0.65	0.01	15405	0.75	0.75	0.00	3390	0.80	0.78	0.02	2980	0.74	0.81	-0.07
AWARDING	Low, Mid	0.65	0.47	0.30	16605	0.57	0.57	0.00	15213	0.70	0.69	0.01	3364	0.78	0.73	0.05	2948	0.75	0.77	-0.01
TRUNK	Low, Mid	0.55	0.42	0.18	16737	0.53	0.51	0.02	15311	0.65	0.64	0.01	3377	0.68	0.68	0.00	2967	0.63	0.72	-0.09
MOM	Low, Mid	1.48	0.53	0.00	16836	0.29	0.29	0.00	15380	0.60	0.59	0.01	3389	0.68	0.69	-0.01	2983	0.81	0.78	0.03
YELLOW	Low, Mid	1.21	0.68	0.00	16832	0.22	0.26	-0.04	15367	0.54	0.52	0.02	3392	0.70	0.61	0.09	2983	0.80	0.70	0.11
YOU	Low, Mid	1.77	1.04	0.00	16809	0.10	0.13	-0.03	15355	0.37	0.36	0.01	3390	0.54	0.47	0.07	2982	0.75	0.59	0.16
BOYBIRD	Low,Mid,High	2.73	1.66	0.19	13255	0.21	0.23	-0.02	14548	0.35	0.36	-0.01	4387	0.48	0.47	0.01	15954	0.84	0.82	0.02
KAYLAFLY	Low,Mid,High	0.40	-0.24	0.00	16928	0.54	0.53	0.01	15403	0.65	0.65	0.00	3391	0.69	0.68	0.01	2981	0.67	0.72	-0.05

BEST COPY AVAILABLE

Table D-1.—Reading item parameters and item fit by rounds (continued)

Reading	Test Form(s)	IRT Parameters			Round 1				Round 2				Round 3				Round 4			
						P+				P+				P+				P+		
		a	b	c	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
COULDNOT	Low,Mid,High	0.54	-0.34	0.00	16895	0.57	0.56	0.01	15381	0.70	0.71	0.00	3389	0.77	0.75	0.02	2978	0.73	0.79	-0.05
COULD	Low,Mid,High	0.37	0.45	0.00	16884	0.44	0.43	0.01	15367	0.53	0.54	-0.01	3389	0.58	0.57	0.01	2982	0.58	0.60	-0.02
BEG WORD	Mid, High	0.62	0.80	0.01	4274	0.64	0.58	0.06	12413	0.65	0.63	0.02	3988	0.70	0.68	0.02	15713	0.76	0.80	-0.04
? MARK	Mid, High	0.79	0.85	0.01	4270	0.50	0.58	-0.08	12410	0.65	0.65	0.01	3987	0.76	0.70	0.06	15719	0.83	0.84	0.00
TIME	Mid	0.74	0.30	0.23	3610	0.77	0.76	0.01	8891	0.78	0.79	-0.01	2333	0.82	0.81	0.02	2368	0.82	0.83	-0.01
JOGGING	Mid	0.84	0.55	0.16	3597	0.73	0.69	0.04	8890	0.73	0.73	0.00	2328	0.79	0.76	0.03	2363	0.70	0.79	-0.09
OR SAT	Mid	1.90	1.25	0.00	3606	0.31	0.31	0.00	8883	0.43	0.41	0.03	2329	0.52	0.48	0.05	2365	0.46	0.57	-0.11
OR PIG	Mid	1.45	1.20	0.00	3612	0.41	0.36	0.05	8895	0.46	0.45	0.01	2329	0.49	0.50	-0.02	2365	0.47	0.58	-0.10
OR TAIL	Mid	2.16	1.38	0.00	3611	0.28	0.21	0.06	8888	0.33	0.30	0.03	2329	0.36	0.37	-0.01	2361	0.38	0.48	-0.09
OR HAND	Mid	2.22	1.48	0.00	3608	0.21	0.16	0.05	8888	0.28	0.24	0.04	2329	0.31	0.30	0.01	2363	0.28	0.40	-0.12
CATCH	Mid, High	2.69	1.75	0.00	4248	0.14	0.13	0.01	12369	0.22	0.23	0.00	3983	0.31	0.34	-0.03	15714	0.77	0.75	0.02
FISHING	Mid, High	3.57	1.72	0.00	4239	0.12	0.12	0.00	12367	0.20	0.22	-0.02	3985	0.31	0.34	-0.03	15713	0.82	0.78	0.04
CANINBAG	Mid, High	1.58	1.74	0.23	2943	0.41	0.41	0.00	9590	0.49	0.48	0.01	3488	0.59	0.55	0.04	15390	0.79	0.78	0.01
KITN BED	Mid, High	2.37	1.74	0.17	3001	0.32	0.32	0.00	9162	0.41	0.41	0.00	3291	0.52	0.50	0.01	15261	0.82	0.80	0.02
GIRLREAD	Mid, High	2.27	2.05	0.25	3068	0.34	0.33	0.01	9647	0.36	0.37	-0.01	3482	0.45	0.44	0.01	15399	0.70	0.69	0.01
KIM CAT	Mid	3.40	1.99	0.51	894	0.52	0.51	0.01	1939	0.54	0.52	0.02	598	0.54	0.52	0.02	1149	0.57	0.54	0.02
NEEDHOME	Mid	4.02	1.56	0.17	796	0.23	0.24	-0.01	1790	0.29	0.30	-0.01	570	0.37	0.36	0.01	1139	0.61	0.50	0.11
LIKE DRY	Mid	3.04	2.16	0.25	784	0.31	0.26	0.05	1747	0.27	0.26	0.00	551	0.23	0.27	-0.03	1093	0.28	0.28	0.00
LIGHT	High	4.64	1.99	0.00	654	0.47	0.42	0.06	3515	0.38	0.38	0.00	1655	0.45	0.47	-0.02	13347	0.75	0.75	0.00
KNOW	High	1.98	1.95	0.00	654	0.45	0.48	-0.03	3512	0.38	0.45	-0.07	1655	0.47	0.52	-0.05	13348	0.74	0.71	0.03
ELEPHANT	High	2.88	1.99	0.00	652	0.48	0.44	0.04	3505	0.41	0.40	0.01	1653	0.49	0.49	0.00	13348	0.72	0.72	0.00
WRONG	High	2.73	2.20	0.00	653	0.37	0.32	0.05	3510	0.34	0.28	0.05	1652	0.34	0.36	-0.03	13346	0.57	0.58	-0.01
ENVELOPE	High	2.78	2.48	0.00	649	0.28	0.20	0.08	3502	0.20	0.16	0.04	1654	0.24	0.22	0.02	13338	0.37	0.39	-0.01
THROUGH	High	2.64	2.49	0.00	653	0.25	0.20	0.05	3504	0.17	0.16	0.01	1651	0.23	0.22	0.01	13346	0.38	0.38	0.00
RAGE	High	2.61	2.67	0.00	653	0.19	0.13	0.06	3510	0.12	0.11	0.02	1654	0.15	0.15	-0.01	13348	0.27	0.27	0.00
TOIL	High	1.82	2.75	0.00	653	0.21	0.14	0.06	3509	0.14	0.12	0.02	1655	0.17	0.16	0.01	13344	0.26	0.27	-0.01
DOGHOUSE	High	2.10	2.47	0.17	532	0.40	0.38	0.02	2569	0.35	0.36	-0.01	1301	0.42	0.41	0.01	12098	0.53	0.53	0.01
FLATTIRE	High	2.71	2.17	0.16	563	0.44	0.48	-0.04	2817	0.43	0.45	-0.03	1438	0.47	0.51	-0.04	12856	0.70	0.68	0.02
MARCHED	High	3.54	2.68	0.20	524	0.36	0.32	0.05	2531	0.31	0.29	0.01	1266	0.31	0.34	-0.03	11576	0.42	0.43	0.00
CHOCCAKE	High	3.89	2.34	0.17	530	0.45	0.41	0.05	2582	0.37	0.37	0.00	1311	0.42	0.44	-0.02	12163	0.59	0.59	0.00
RECIPE	High	2.79	2.93	0.19	509	0.34	0.26	0.08	2403	0.30	0.25	0.05	1223	0.30	0.28	0.02	11750	0.30	0.32	-0.03
INGREDNT	High	3.51	3.03	0.19	506	0.22	0.22	0.00	2397	0.25	0.22	0.03	1238	0.25	0.24	0.01	11410	0.25	0.27	-0.02
CAPTURE	High	2.16	2.81	0.00	643	0.15	0.11	0.04	3466	0.10	0.09	0.01	1644	0.11	0.13	-0.02	13287	0.22	0.22	0.00
CORNER	High	1.87	2.78	0.00	645	0.17	0.13	0.04	3466	0.12	0.11	0.01	1644	0.15	0.15	0.00	13282	0.25	0.25	0.00
WEB	High	1.48	2.86	0.00	645	0.18	0.14	0.04	3464	0.12	0.12	0.00	1643	0.15	0.16	-0.01	13273	0.26	0.25	0.00
STRANDS	High	1.35	3.38	0.00	645	0.08	0.06	0.02	3461	0.06	0.05	0.01	1643	0.06	0.07	-0.01	13268	0.11	0.11	0.00
CAT NAME	High suppl.	1.94	2.68	0.00									1506	0.22	0.19	0.03	13309	0.30	0.30	0.00

Table D-1.—Reading item parameters and item fit by rounds (continued)

Reading	Test Form(s)	IRT Parameters			Round 1				Round 2				Round 3				Round 4			
					N	P+		Difference	N	P+		Difference	N	P+		Difference	N	P+		Difference
		a	b	c		Actual	Predicted			Actual	Predicted			Actual	Predicted			Actual	Predicted	
OWNRNAME	High suppl.	1.92	2.78	0.00									1525	0.16	0.16	0.00	13314	0.25	0.25	0.00
APPROX	High suppl.	1.87	3.18	0.00									1519	0.07	0.07	0.00	13310	0.11	0.11	0.00
MOREINFO	High suppl.	1.41	3.15	0.00									1500	0.09	0.11	-0.02	13306	0.17	0.16	0.00
QUIET	High suppl.	2.81	2.44	0.00									1509	0.26	0.25	0.01	13334	0.41	0.41	0.00
WTLESS	High suppl.	4.27	2.72	0.00									1487	0.13	0.12	0.01	13324	0.22	0.22	0.00
REQUIRE	High suppl.	3.51	2.78	0.00									1456	0.14	0.11	0.02	13330	0.19	0.19	0.00
UNUSUAL	High suppl.	3.50	2.87	0.00									548	0.28	0.22	0.05	6180	0.30	0.31	0.00
MOISTURE	High suppl.	2.47	2.92	0.00									358	0.32	0.30	0.02	3929	0.40	0.41	0.00
WAGES	High suppl.	1.42	3.48	0.00									308	0.18	0.15	0.03	3263	0.20	0.20	0.00
VICIOUS	High suppl.	2.62	3.29	0.00									257	0.18	0.13	0.05	2509	0.20	0.21	0.00
PREFRNCE	High suppl.	1.17	3.33	0.00									220	0.28	0.24	0.04	1858	0.33	0.34	0.00
AMBITIO	High suppl.	2.07	3.46	0.00									179	0.11	0.12	0.00	1207	0.22	0.22	0.00
CRITCISM	High suppl.	1.71	3.75	0.00									158	0.11	0.07	0.04	905	0.15	0.16	-0.01
MYSTERLY	High suppl.	2.47	3.15	0.00									147	0.22	0.21	0.01	727	0.44	0.45	0.00
ALIGNMNT	High suppl.	0.75	4.95	0.00									129	0.05	0.05	-0.01	438	0.11	0.10	0.01
MAKE $	High suppl.	1.00	1.93	0.23									187	0.79	0.74	0.04	356	0.91	0.92	-0.01
MAINIDEA	High suppl.	1.39	3.28	0.30									154	0.47	0.43	0.04	331	0.66	0.64	0.02
WHY NO $	High suppl.	1.70	3.33	0.19									161	0.24	0.31	-0.06	329	0.60	0.56	0.04
DESCRIBE	High suppl.	1.90	3.68	0.13									170	0.15	0.18	-0.03	351	0.34	0.34	0.00

•Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

Table D-2.—Mathematics item parameters and item fit by rounds

					Round 1				Round 2				Round 3*				Round 4			
		IRT Parameters				P+				P+				P+				P+		
Mathematics	Test Form(s)	a	b	c	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
SM-LG-SM	Routing	0.82	0.05	0.27	18379	0.60	0.60	0.00	19471	0.80	0.78	0.01	5206	0.89	0.86	0.03	16587	0.94	0.95	-0.01
COUNT20	Routing	0.68	-0.44	0.00	18622	0.56	0.57	-0.01	19651	0.80	0.77	0.03	5222	0.85	0.85	0.00	16645	0.91	0.94	-0.02
COUNT10*	Routing	0.48	-3.10	0.01	18622	0.90	0.90	0.00	19651	0.96	0.95	0.01	5222	0.96	0.97	0.00	16645	0.97	0.98	-0.01
NUMBER9	Routing	1.35	-0.69	0.00	18631	0.66	0.67	-0.01	19652	0.90	0.90	0.01	5226	0.95	0.95	0.00	16647	0.99	0.99	0.00
NUMBER23	Routing	1.14	0.43	0.00	18614	0.33	0.34	-0.01	19639	0.65	0.64	0.01	5225	0.77	0.77	-0.01	16647	0.94	0.93	0.01
3RD LINE	Routing	1.11	0.50	0.01	18636	0.33	0.32	0.01	19653	0.63	0.62	0.01	5225	0.76	0.75	0.00	16647	0.91	0.92	-0.01
STICKBAT	Routing	0.54	-1.02	0.13	18616	0.74	0.71	0.03	19647	0.84	0.84	-0.01	5224	0.89	0.89	-0.01	16643	0.92	0.95	-0.03
_789 10	Routing	1.12	0.58	0.02	18621	0.28	0.31	-0.04	19649	0.63	0.60	0.02	5225	0.78	0.74	0.04	16644	0.93	0.92	0.01
51015_25	Routing	1.29	1.93	0.00	18618	0.04	0.06	-0.02	19636	0.19	0.19	-0.01	5224	0.30	0.33	-0.03	16644	0.71	0.66	0.05
3+2 CARS	Routing	0.93	0.75	0.11	18636	0.38	0.35	0.03	19656	0.59	0.59	0.00	5225	0.71	0.71	0.00	16646	0.86	0.89	-0.03
5-1ORANG	Routing	1.12	1.41	0.13	18427	0.27	0.24	0.03	19524	0.43	0.43	0.01	5215	0.56	0.56	0.00	16585	0.79	0.81	-0.02
2+5MARBL	Routing	0.87	1.35	0.00	18616	0.19	0.17	0.02	19645	0.38	0.37	0.00	5223	0.53	0.51	0.02	16645	0.75	0.76	-0.01
3+7PENNY	Routing	1.35	1.77	0.03	18623	0.10	0.09	0.01	19645	0.24	0.24	0.00	5224	0.38	0.39	-0.01	16645	0.72	0.72	0.01
13___79	Routing	1.21	2.52	0.00	4259	0.09	0.09	0.00	11212	0.14	0.16	-0.02	3872	0.21	0.24	-0.03	15550	0.50	0.47	0.03
COST$10	Routing	1.71	2.69	0.04	4173	0.11	0.08	0.03	11191	0.14	0.12	0.02	3866	0.22	0.18	0.04	15546	0.39	0.42	-0.02
8/2CANDY	Routing	1.56	3.03	0.02	4260	0.05	0.04	0.01	11204	0.08	0.07	0.01	3872	0.15	0.11	0.04	15548	0.26	0.27	-0.01
15/5CARS	Routing	1.98	2.80	0.04	4260	0.08	0.07	0.01	11201	0.11	0.10	0.01	3871	0.17	0.15	0.02	15542	0.35	0.36	-0.01
2CRAYONS	Low	0.86	-3.73	0.02	14378	0.98	0.98	0.00	8443	0.99	0.99	0.00	1353	0.99	0.99	-0.01	1097	0.98	1.00	-0.01
3BANANAS	Low	0.44	-3.12	0.11	14289	0.89	0.88	0.01	8412	0.91	0.91	-0.01	1350	0.92	0.92	-0.01	1092	0.87	0.93	-0.07
6BANANAS	Low	0.62	-0.25	0.01	14364	0.44	0.45	-0.01	8440	0.59	0.57	0.02	1352	0.66	0.61	0.04	1095	0.71	0.66	0.05
NUMBER 4	Low	1.83	-1.65	0.06	14369	0.87	0.88	-0.01	8441	0.96	0.97	-0.01	1353	0.96	0.98	-0.01	1097	0.97	0.98	-0.01
NUMBER 7	Low	1.57	-1.19	0.01	14372	0.74	0.76	-0.01	8441	0.91	0.90	0.00	1353	0.90	0.93	-0.02	1096	0.93	0.95	-0.02
NUMBER17	Low, Mid	1.10	0.14	0.00	17477	0.37	0.39	-0.02	14603	0.66	0.63	0.03	2873	0.71	0.71	-0.01	3411	0.84	0.82	0.02
SQUARE	Low	0.52	-2.72	0.20	14356	0.90	0.88	0.01	8427	0.91	0.93	-0.02	1350	0.90	0.94	-0.04	1097	0.90	0.95	-0.05
LG-SM-SM	Low, Mid	0.87	-0.02	0.30	17287	0.61	0.61	0.00	14487	0.77	0.76	0.01	2862	0.83	0.81	0.02	3376	0.87	0.87	0.00
000X	Low, Mid	0.64	0.33	0.21	16857	0.51	0.51	0.00	14133	0.67	0.64	0.03	2823	0.75	0.69	0.06	3327	0.76	0.76	0.00
HALFOVAL	Low, Mid	0.56	0.63	0.25	16822	0.50	0.49	0.00	14113	0.63	0.60	0.03	2798	0.68	0.65	0.03	3320	0.73	0.71	0.02
2+3STICK	Low,Mid,High	1.09	0.87	0.10	18623	0.32	0.31	0.01	19650	0.56	0.55	0.00	5224	0.71	0.69	0.01	16645	0.88	0.89	-0.01
3-1PENCL	Low, Mid	0.47	-1.13	0.05	17494	0.68	0.67	0.01	14609	0.78	0.78	0.00	2873	0.80	0.81	-0.01	3411	0.80	0.85	-0.05
2+5CIRCL	Low,Mid,High	0.95	1.59	0.01	18623	0.15	0.13	0.02	19648	0.32	0.31	0.01	5221	0.46	0.45	0.01	16643	0.70	0.72	-0.02
8-6CRAYN	Low,Mid,High	0.93	1.37	0.08	18625	0.22	0.22	0.00	19651	0.42	0.41	0.01	5222	0.53	0.55	-0.01	16644	0.80	0.78	0.01
PNTBRUSH	Low,Mid,High	0.91	-0.45	0.24	18299	0.68	0.69	-0.01	19480	0.87	0.86	0.01	5200	0.93	0.92	0.01	16622	0.98	0.98	0.00
#CHOC	Low,Mid,High	0.74	-0.71	0.01	18609	0.64	0.63	0.00	19647	0.83	0.83	0.00	5222	0.90	0.89	0.01	16643	0.96	0.96	0.00
#VANILLA	Low,Mid,High	0.70	-0.99	0.04	18608	0.70	0.70	0.00	19647	0.86	0.87	0.00	5222	0.93	0.92	0.02	16643	0.97	0.97	0.00
#BUGS	Low,Mid,High	0.86	0.75	0.23	18166	0.48	0.45	0.03	19399	0.65	0.64	0.00	5166	0.75	0.75	0.00	16564	0.88	0.90	-0.02
4 LINES	Mid	0.35	-0.03	0.26	3032	0.76	0.73	0.03	6041	0.77	0.75	0.02	1499	0.80	0.76	0.04	2275	0.73	0.79	-0.06

Table D-2.—Mathematics item parameters and item fit by rounds (continued)

Mathematics	Test Form(s)	IRT Parameters			Round 1				Round 2				Round 3				Round 4			
					N	P+		Difference	N	P+		Difference	N	P+		Difference	N	P+		Difference
		a	b	c		Actual	Predicted			Actual	Predicted			Actual	Predicted			Actual	Predicted	
SHAPES	Mid	0.46	1.94	0.30	3033	0.52	0.52	0.00	6049	0.55	0.54	0.01	1505	0.60	0.56	0.04	2294	0.59	0.59	0.00
PATTERN	Mid, High	0.85	1.63	0.25	4243	0.50	0.51	-0.01	11185	0.59	0.59	0.00	3867	0.67	0.65	0.02	15533	0.79	0.80	-0.01
12 BY 2S	Mid, High	1.16	1.62	0.00	4253	0.31	0.31	0.00	11199	0.43	0.44	-0.01	3863	0.52	0.55	-0.04	15546	0.79	0.78	0.01
2+2	Mid	3.63	0.90	0.04	3122	0.51	0.53	-0.02	6167	0.67	0.69	-0.02	1520	0.77	0.79	-0.02	2317	0.89	0.91	-0.01
3+4	Mid, High	1.32	1.36	0.00	4258	0.34	0.39	-0.06	11200	0.54	0.54	0.00	3869	0.67	0.66	0.01	15547	0.86	0.86	0.00
1+7	Mid	3.26	1.56	0.42	3121	0.48	0.47	0.01	6159	0.53	0.52	0.01	1518	0.62	0.56	0.06	2315	0.72	0.68	0.04
3+3	Mid	4.29	1.00	0.01	3120	0.43	0.43	0.00	6165	0.60	0.62	-0.01	1519	0.71	0.73	-0.02	2317	0.86	0.88	-0.02
11+3	Mid, High	1.30	1.95	0.00	4250	0.17	0.19	-0.02	11179	0.30	0.31	-0.01	3866	0.43	0.42	0.01	15542	0.70	0.69	0.01
12+6	Mid, High	1.09	2.35	0.00	4243	0.12	0.13	-0.01	11169	0.21	0.21	0.00	3860	0.31	0.30	0.01	15539	0.54	0.53	0.01
17-4	Mid, High	1.52	2.64	0.00	4245	0.04	0.05	-0.01	11168	0.09	0.10	-0.01	3859	0.15	0.17	-0.02	15536	0.44	0.42	0.03
#STRAW	High	0.67	-1.23	0.05	1136	0.95	0.97	-0.02	5042	0.97	0.97	0.00	2351	0.98	0.98	0.01	13232	0.99	0.99	0.00
#MORE	High	2.09	2.66	0.20	1135	0.35	0.28	0.07	5038	0.36	0.32	0.04	2351	0.43	0.37	0.05	13221	0.54	0.57	-0.03
HEADSUP	High	0.90	3.27	0.09	1136	0.23	0.19	0.04	5040	0.22	0.21	0.00	2351	0.26	0.25	0.01	13225	0.35	0.35	-0.01
HOWMANY$	High	1.39	3.15	0.11	1135	0.19	0.17	0.03	5035	0.20	0.19	0.01	2351	0.25	0.22	0.02	13229	0.34	0.35	-0.01
25-14BKS	High	1.72	3.08	0.00	1136	0.07	0.05	0.02	5038	0.09	0.08	0.01	2350	0.13	0.12	0.01	13229	0.26	0.27	-0.01
12-? PEN	High	1.88	3.08	0.03	1136	0.12	0.08	0.05	5040	0.13	0.10	0.03	2351	0.16	0.14	0.02	13226	0.27	0.28	-0.02
GOALS	High	1.72	3.19	0.01	1135	0.05	0.04	0.01	5036	0.07	0.06	0.01	2351	0.10	0.10	0.01	13226	0.23	0.23	0.00
CHANGE	High	1.34	4.35	0.00	1135	0.04	0.01	0.04	5037	0.01	0.01	0.00	2351	0.02	0.02	0.00	13229	0.03	0.04	0.00
17CENTS	High	1.98	3.33	0.01	1134	0.02	0.03	-0.01	5030	0.04	0.04	0.00	2350	0.05	0.06	-0.01	13221	0.17	0.17	0.01
BD CAKE	High	1.43	3.40	0.01	1135	0.04	0.04	-0.01	5036	0.05	0.06	-0.01	2347	0.07	0.08	-0.01	13223	0.19	0.19	0.01
24/4 TAB	High	1.80	3.59	0.04	1135	0.06	0.06	0.01	5034	0.08	0.06	0.01	2347	0.09	0.08	0.01	13217	0.14	0.14	-0.01
2-1+2	High	1.42	2.62	0.13	1134	0.28	0.26	0.02	5027	0.30	0.31	-0.01	2347	0.35	0.37	-0.01	13221	0.56	0.55	0.01
9-2	High	1.68	2.05	0.00	1135	0.25	0.33	-0.08	5035	0.35	0.43	-0.07	2347	0.43	0.52	-0.09	13230	0.80	0.76	0.04
7-3	High	1.60	2.01	0.01	1136	0.27	0.35	-0.08	5034	0.38	0.45	-0.07	2347	0.45	0.54	-0.09	13229	0.81	0.77	0.04
4+4-2	High	1.42	2.58	0.01	1131	0.16	0.17	-0.01	5025	0.23	0.23	0.00	2345	0.30	0.30	0.01	13222	0.51	0.51	0.00
6+7	High	1.34	2.19	0.00	1130	0.28	0.30	-0.02	5022	0.36	0.38	-0.02	2346	0.45	0.46	-0.01	13225	0.68	0.68	0.00
12-9	High	1.51	2.67	0.00	1131	0.10	0.13	-0.03	5020	0.14	0.19	-0.04	2343	0.20	0.25	-0.05	13218	0.49	0.46	0.03
26+20	High	1.58	2.76	0.00	1131	0.10	0.11	-0.01	5000	0.13	0.15	-0.03	2339	0.17	0.21	-0.04	13210	0.44	0.42	0.02

*Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

Table D-3.—General knowledge item parameters and item fit by rounds

General knowledge	Test Form(s)	IRT Parameters			Round 1				Round 2				Round 3*				Round 4			
					N	P+		Difference	N	P+		Difference	N	P+		Difference	N	P+		Difference
		a	b	c		Actual	Predicted			Actual	Predicted			Actual	Predicted			Actual	Predicted	
WHY SCHL	Routing	1.58	-0.83	0.03	17551	0.72	0.73	-0.01	18898	0.84	0.85	0.00	5043	0.90	0.90	0.00	16323	0.95	0.94	0.00
BEEHONEY	Routing	1.32	-0.63	0.00	17550	0.64	0.63	0.01	18880	0.75	0.76	-0.01	5043	0.82	0.82	0.00	16323	0.89	0.89	0.00
STRTFIRE	Routing	1.28	-0.55	0.05	17524	0.57	0.62	-0.04	18881	0.76	0.74	0.01	5036	0.85	0.81	0.05	16311	0.89	0.88	0.01
THERMOM	Routing	1.41	-0.17	0.00	17551	0.45	0.44	0.01	18894	0.59	0.59	-0.01	5042	0.66	0.68	-0.03	16325	0.80	0.79	0.01
WASHNGTN	Routing	1.62	0.14	0.00	17512	0.28	0.30	-0.02	18871	0.47	0.46	0.02	5040	0.54	0.56	-0.03	16317	0.72	0.70	0.02
AIRLUNGS	Routing	1.10	0.00	0.02	17534	0.43	0.40	0.03	18872	0.52	0.53	-0.01	5037	0.60	0.61	-0.01	16319	0.70	0.71	-0.01
HEALTHY	Routing	1.32	-0.25	0.00	17568	0.46	0.47	-0.01	18905	0.63	0.62	0.00	5044	0.71	0.71	0.00	16327	0.80	0.80	0.00
ARTIST	Routing	1.82	0.01	0.00	17532	0.37	0.34	0.03	18869	0.51	0.52	-0.01	5040	0.64	0.63	0.01	16319	0.74	0.76	-0.02
HEART	Routing	1.23	0.19	0.00	17528	0.30	0.30	-0.01	18877	0.44	0.44	0.00	5037	0.55	0.53	0.02	16319	0.66	0.65	0.01
FROG	Routing	1.74	0.02	0.01	17520	0.37	0.35	0.02	18874	0.52	0.52	0.00	5040	0.63	0.62	0.01	16318	0.73	0.75	-0.02
FARMER	Routing	1.59	0.84	0.00	17509	0.11	0.09	0.03	18863	0.18	0.17	0.01	5032	0.25	0.25	0.00	16309	0.35	0.38	-0.02
WHEAT	Routing	1.55	1.06	0.00	17498	0.07	0.06	0.01	18851	0.12	0.12	0.01	5033	0.18	0.18	0.00	16297	0.28	0.28	-0.01
CANDLE	Low	0.85	-1.73	0.15	12213	0.84	0.85	0.00	9295	0.88	0.88	0.00	1791	0.91	0.89	0.02	3429	0.90	0.90	0.00
CRAB	Low	1.05	-1.50	0.19	12223	0.85	0.85	0.00	9293	0.88	0.88	0.00	1788	0.90	0.90	0.00	3427	0.90	0.91	-0.01
GLOBE	Low	0.99	-1.67	0.11	12242	0.86	0.86	0.00	9301	0.89	0.89	0.00	1789	0.91	0.90	0.01	3424	0.91	0.92	-0.01
MAGNET	Low	0.81	-1.23	0.13	12107	0.73	0.73	-0.01	9221	0.78	0.78	0.00	1779	0.82	0.80	0.02	3411	0.85	0.82	0.04
EARRADIO	Low	0.92	-0.45	0.44	12088	0.67	0.69	-0.02	9222	0.76	0.73	0.03	1779	0.78	0.74	0.04	3400	0.80	0.77	0.03
SINK	Low	0.77	-1.17	0.12	12188	0.69	0.71	-0.02	9285	0.77	0.76	0.02	1786	0.81	0.77	0.03	3426	0.82	0.79	0.03
WINTER	Low, High	0.85	-0.94	0.03	17428	0.68	0.70	-0.01	18833	0.80	0.78	0.01	5025	0.84	0.83	0.02	16294	0.87	0.88	0.00
CALL 911	Low	1.46	-0.74	0.01	12247	0.59	0.59	0.01	9305	0.67	0.68	-0.01	1790	0.71	0.71	0.00	3432	0.74	0.76	-0.02
ICE CUBE	Low, High	1.52	-0.82	0.00	17554	0.72	0.72	0.00	18886	0.84	0.84	0.01	5043	0.89	0.89	0.00	16324	0.92	0.94	-0.02
WATERMAP	Low	0.61	-0.40	0.01	12253	0.44	0.46	-0.02	9304	0.51	0.50	0.01	1791	0.55	0.52	0.03	3433	0.59	0.55	0.04
S POLE	Low, High	1.55	-0.47	0.01	17525	0.60	0.57	0.03	18875	0.73	0.72	0.00	5037	0.79	0.80	-0.01	16320	0.85	0.88	-0.03
CLOUDS	Low	1.48	-0.42	0.00	12256	0.46	0.42	0.03	9310	0.50	0.51	-0.01	1792	0.53	0.56	-0.02	3426	0.55	0.61	-0.06
RULES	Low, High	0.94	-0.12	0.00	17538	0.48	0.43	0.04	18875	0.54	0.55	-0.01	5037	0.61	0.63	-0.02	16318	0.70	0.72	-0.02
GARDEN	Low	1.83	-1.09	0.00	12279	0.76	0.77	-0.01	9321	0.84	0.85	-0.01	1794	0.86	0.88	-0.01	3436	0.89	0.90	-0.01
HELPBODY	Low, High	1.05	-0.54	0.00	17561	0.56	0.58	-0.02	18910	0.70	0.70	0.00	5044	0.77	0.77	0.00	16327	0.86	0.84	0.02
DANGER	Low, High	1.10	-0.05	0.00	17564	0.36	0.40	-0.04	18908	0.54	0.54	0.00	5044	0.66	0.62	0.04	16328	0.74	0.72	0.02
PILGRIMS	Low, High	1.54	0.23	0.07	17555	0.28	0.32	-0.03	18902	0.47	0.46	0.02	5042	0.56	0.55	0.00	16323	0.70	0.68	0.02
BAD AIR	Low, High	1.75	0.49	0.22	17557	0.37	0.35	0.02	18906	0.45	0.45	0.01	5042	0.53	0.53	0.00	16326	0.63	0.64	-0.02
SUMMER	Low, High	0.88	0.42	0.00	17530	0.22	0.28	-0.06	18881	0.44	0.38	0.07	5040	0.36	0.45	-0.09	16318	0.57	0.55	0.02

Table D-3.—General knowledge item parameters and item fit by rounds (continued)

General knowledge	Test Form(s)	IRT Parameters			Round 1				Round 2				Round 3				Round 4			
					N	P+		Difference	N	P+		Difference	N	P+		Difference	N	P+		Difference
		a	b	c		Actual	Predicted			Actual	Predicted			Actual	Predicted			Actual	Predicted	
POLLUTE	Low, High	1.17	0.11	0.00	17509	0.39	0.34	0.05	18873	0.47	0.47	-0.01	5035	0.53	0.56	-0.03	16313	0.65	0.68	-0.03
WHY LAWS	Low, High	1.22	0.55	0.00	17485	0.23	0.19	0.03	18842	0.31	0.30	0.00	5030	0.37	0.39	-0.02	16310	0.49	0.51	-0.02
TRAV OLD	Low, High	2.15	0.55	0.14	17004	0.26	0.25	0.01	18579	0.36	0.35	0.00	5008	0.45	0.45	0.01	16216	0.59	0.59	0.00
BIRD EAT	Low, High	0.67	0.37	0.14	17123	0.43	0.43	0.01	18678	0.50	0.50	0.00	5005	0.57	0.55	0.02	16217	0.61	0.62	-0.01
OWL EAT	Low, High	1.01	0.34	0.24	16420	0.46	0.45	0.01	17983	0.56	0.54	0.02	4889	0.62	0.60	0.02	15953	0.69	0.68	0.01
LEADERS	Low, High	1.71	1.32	0.00	17470	0.03	0.03	0.01	18829	0.06	0.06	0.00	5022	0.09	0.10	0.00	16290	0.19	0.18	0.01
BALANCE	High	0.98	0.01	0.23	5252	0.67	0.70	-0.03	9527	0.73	0.75	-0.02	3245	0.80	0.77	0.03	12861	0.83	0.82	0.02
MICRSCOP	High	1.02	0.52	0.18	5189	0.44	0.51	-0.07	9395	0.56	0.57	0.00	3217	0.64	0.61	0.03	12721	0.71	0.67	0.04
FLATTIRE	High	0.64	0.84	0.32	5159	0.56	0.57	-0.01	9381	0.59	0.59	0.00	3208	0.61	0.61	-0.01	12639	0.67	0.65	0.02
EAT STOM	High	1.25	1.05	0.25	5126	0.42	0.39	0.03	9275	0.45	0.43	0.02	3183	0.50	0.47	0.03	12548	0.52	0.53	-0.01
TURTLE	High	0.65	2.10	0.31	5134	0.41	0.40	0.02	9322	0.43	0.41	0.02	3190	0.43	0.42	0.01	12554	0.44	0.45	-0.01
INDIANS	High	1.58	0.59	0.05	5276	0.41	0.37	0.05	9578	0.47	0.45	0.02	3248	0.47	0.52	-0.05	12886	0.60	0.61	-0.02
COINFACE	High	1.13	0.47	0.00	5279	0.36	0.42	-0.06	9577	0.51	0.49	0.02	3246	0.53	0.55	-0.02	12886	0.64	0.62	0.02
ROOTS	High	1.36	0.72	0.00	5282	0.33	0.29	0.03	9577	0.38	0.37	0.01	3249	0.44	0.43	0.01	12889	0.51	0.52	-0.01
US ONMAP	High	1.02	0.61	0.27	5011	0.53	0.54	-0.02	8998	0.61	0.59	0.02	3072	0.64	0.62	0.01	12282	0.73	0.68	0.05
NORTH US	High	1.93	0.84	0.44	4969	0.50	0.54	-0.04	9087	0.58	0.59	-0.01	3154	0.65	0.63	0.02	12542	0.76	0.69	0.06
SOUTH US	High	1.83	0.96	0.33	5005	0.40	0.43	-0.03	9097	0.47	0.48	-0.01	3164	0.54	0.52	0.02	12563	0.64	0.59	0.05
ISLAND	High	1.76	1.12	0.00	5274	0.14	0.10	0.04	9564	0.17	0.16	0.01	3247	0.24	0.21	0.03	12888	0.29	0.31	-0.01
ML KING	High	1.42	1.66	0.01	5272	0.02	0.06	-0.04	9562	0.08	0.08	0.00	3247	0.08	0.10	-0.03	12881	0.19	0.15	0.04
JULY 4	High	1.51	1.89	0.02	5279	0.06	0.04	0.02	9567	0.06	0.05	0.00	3242	0.08	0.07	0.01	12884	0.09	0.10	-0.01

*Round 3 is a subset of approximately 30 percent of the full ECLS–K sample.

Table D-4.—Mathematics item parameters and comparison of item fit for English versus Spanish version, all rounds combined

		IRT Parameters			English Version				Spanish Version			
						P+				P+		
Math	Test Form(s)	a	b	c	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
SM-LG-SM	Routing	0.82	0.05	0.27	57458	0.79	0.79	0.01	2185	0.53	0.56	-0.03
COUNT20	Routing	0.68	-0.44	0.00	57912	0.78	0.77	0.00	2228	0.42	0.51	-0.10
COUNT10*	Routing	0.48	-3.10	0.01	57912	0.95	0.95	0.00	2228	0.80	0.87	-0.07
NUMBER9	Routing	1.35	-0.69	0.00	57928	0.87	0.87	0.00	2228	0.59	0.57	0.02
NUMBER23	Routing	1.14	0.43	0.00	57897	0.66	0.65	0.00	2228	0.28	0.30	-0.02
3RD LINE	Routing	1.11	0.50	0.01	57931	0.64	0.63	0.01	2230	0.16	0.29	-0.13
STICKBAT	Routing	0.54	-1.02	0.13	57904	0.85	0.84	0.01	2226	0.49	0.67	-0.18
_789 10	Routing	1.12	0.58	0.02	57913	0.63	0.62	0.00	2226	0.34	0.29	0.06
51015_25	Routing	1.29	1.93	0.00	57898	0.31	0.30	0.01	2224	0.07	0.09	-0.02
3+2 CARS	Routing	0.93	0.75	0.11	57934	0.62	0.62	0.01	2229	0.29	0.34	-0.05
5-1ORANG	Routing	1.12	1.41	0.13	57560	0.50	0.50	0.00	2191	0.26	0.26	0.00
2+5MARBL	Routing	0.87	1.35	0.00	57902	0.44	0.44	0.00	2227	0.22	0.18	0.04
3+7PENNY	Routing	1.35	1.77	0.03	57911	0.35	0.35	0.00	2226	0.13	0.12	0.01
13___79	Routing	1.21	2.52	0.00	34485	0.30	0.30	0.00	408	0.20	0.24	-0.04
COST$10	Routing	1.71	2.69	0.04	34369	0.26	0.26	0.00	407	0.15	0.19	-0.04
8/2CANDY	Routing	1.56	3.03	0.02	34476	0.17	0.16	0.00	408	0.12	0.11	0.01
15/5CARS	Routing	1.98	2.80	0.04	34466	0.22	0.22	0.00	408	0.15	0.16	-0.01
2CRAYONS	Low	0.86	-3.73	0.02	23449	0.99	0.99	0.00	1822	0.96	0.97	-0.01
3BANANAS	Low	0.44	-3.12	0.11	23342	0.90	0.90	0.00	1801	0.82	0.85	-0.03
6BANANAS	Low	0.62	-0.25	0.01	23431	0.52	0.52	0.01	1820	0.41	0.39	0.02
NUMBER 4	Low	1.83	-1.65	0.06	23440	0.92	0.93	-0.01	1820	0.77	0.80	-0.02
NUMBER 7	Low	1.57	-1.19	0.01	23441	0.83	0.84	-0.01	1821	0.62	0.65	-0.03
NUMBER17	Low, Mid	1.10	0.14	0.00	36327	0.56	0.56	0.01	2037	0.28	0.31	-0.03
SQUARE	Low	0.52	-2.72	0.20	23419	0.91	0.91	0.00	1811	0.83	0.86	-0.03
LG-SM-SM	Low, Mid	0.87	-0.02	0.30	36002	0.72	0.71	0.01	2010	0.55	0.55	0.00
000X	Low, Mid	0.64	0.33	0.21	35174	0.62	0.60	0.02	1966	0.46	0.46	0.00
HALFOVAL	Low, Mid	0.56	0.63	0.25	35078	0.59	0.57	0.02	1975	0.44	0.46	-0.02
2+3STICK	Low,Mid,High	1.09	0.87	0.10	57914	0.60	0.59	0.00	2228	0.33	0.30	0.03
3-1PENCL	Low, Mid	0.47	-1.13	0.05	36346	0.75	0.74	0.00	2041	0.55	0.61	-0.06
2+5CIRCL	Low,Mid,High	0.95	1.59	0.01	57907	0.39	0.39	0.00	2228	0.13	0.14	-0.01
8-6CRAYN	Low,Mid,High	0.93	1.37	0.08	57915	0.48	0.48	0.00	2227	0.28	0.23	0.04
PNTBRUSH	Low,Mid,High	0.91	-0.45	0.24	57437	0.86	0.85	0.00	2164	0.65	0.63	0.02
#CHOC	Low,Mid,High	0.74	-0.71	0.01	57892	0.82	0.82	0.00	2229	0.59	0.57	0.02
#VANILLA	Low,Mid,High	0.70	-0.99	0.04	57892	0.85	0.86	0.00	2228	0.67	0.64	0.03
#BUGS	Low,Mid,High	0.86	0.75	0.23	57119	0.68	0.67	0.00	2176	0.43	0.44	0.00

Table D-4.—Mathematics item parameters and comparison of item fit for English versus Spanish version, all rounds combined (continued)

Math	Test Form(s)	IRT Parameters			English Version				Spanish Version			
						P+				P+		
		a	b	c	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
4 LINES	Mid	0.35	-0.03	0.26	12635	0.77	0.75	0.01	212	0.71	0.76	-0.05
SHAPES	Mid	0.46	1.94	0.30	12667	0.55	0.55	0.01	214	0.58	0.56	0.03
PATTERN	Mid, High	0.85	1.63	0.25	34424	0.68	0.68	0.00	404	0.57	0.65	-0.08
12 BY 2S	Mid, High	1.16	1.62	0.00	34454	0.59	0.59	0.00	407	0.58	0.54	0.04
2+2	Mid	3.63	0.90	0.04	12906	0.68	0.70	-0.02	220	0.77	0.76	0.01
3+4	Mid, High	1.32	1.36	0.00	34466	0.67	0.68	0.00	408	0.73	0.64	0.09
1+7	Mid	3.26	1.56	0.42	12893	0.56	0.54	0.02	220	0.70	0.58	0.12
3+3	Mid	4.29	1.00	0.01	12901	0.62	0.63	-0.01	220	0.72	0.71	0.01
11+3	Mid, High	1.30	1.95	0.00	34429	0.48	0.48	0.00	408	0.56	0.42	0.14
12+6	Mid, High	1.09	2.35	0.00	34403	0.36	0.35	0.00	408	0.42	0.30	0.12
17-4	Mid, High	1.52	2.64	0.00	34401	0.25	0.25	0.00	407	0.25	0.18	0.06
#STRAW	High	0.67	-1.23	0.05	21573	0.98	0.98	0.00	188	0.97	0.98	-0.01
#MORE	High	2.09	2.66	0.20	21559	0.48	0.47	0.00	186	0.19	0.45	-0.26
HEADSUP	High	0.90	3.27	0.09	21565	0.30	0.30	0.00	187	0.19	0.28	-0.09
HOWMANY$	High	1.39	3.15	0.11	21563	0.29	0.29	0.00	187	0.27	0.26	0.01
25-14BKS	High	1.72	3.08	0.00	21565	0.20	0.20	0.00	188	0.09	0.15	-0.06
12-? PEN	High	1.88	3.08	0.03	21566	0.22	0.22	0.00	187	0.08	0.17	-0.09
GOALS	High	1.72	3.19	0.01	21560	0.17	0.17	0.00	188	0.15	0.13	0.02
CHANGE	High	1.34	4.35	0.00	21565	0.03	0.03	0.00	187	0.00	0.02	-0.02
17CENTS	High	1.98	3.33	0.01	21548	0.12	0.12	0.00	187	0.09	0.08	0.01
BD CAKE	High	1.43	3.40	0.01	21554	0.14	0.14	0.00	187	0.21	0.11	0.11
24/4 TAB	High	1.80	3.59	0.04	21545	0.11	0.11	0.00	188	0.07	0.08	-0.01
2-1+2	High	1.42	2.62	0.13	21542	0.46	0.46	0.00	187	0.45	0.45	0.00
9-2	High	1.68	2.05	0.00	21559	0.63	0.63	0.00	188	0.71	0.65	0.06
7-3	High	1.60	2.01	0.01	21558	0.64	0.65	-0.01	188	0.78	0.67	0.11
4+4-2	High	1.42	2.58	0.01	21535	0.40	0.40	-0.01	188	0.36	0.39	-0.03
6+7	High	1.34	2.19	0.00	21535	0.56	0.56	0.00	188	0.67	0.57	0.10
12-9	High	1.51	2.67	0.00	21524	0.36	0.36	0.00	188	0.52	0.34	0.19
26+20	High	1.58	2.76	0.00	21492	0.32	0.32	0.00	188	0.40	0.29	0.12

Table D-4.—Mathematics item parameters and comparison of item fit for English versus Spanish version, all rounds combined (continued)

					English Version				Spanish Version			
		IRT Parameters				P+				P+		
Math	Test Form(s)	a	b	c	N	Actual	Predicted	Difference	N	Actual	Predicted	Difference
Avg Difference												
Routing								0.00				-0.03
Low Form								0.00				0.00
Middle Form								0.00				0.02
High Form								0.00				0.02
All Items								0.00				0.00

Table D-5.—Root mean square deviation and mean deviation for English versus Spanish mathematics test items

	Total		English		Spanish	
Item	RMSD	MD	RMSD	MD	RMSD	MD
SM-LG-SM	0.02	0.00	0.02	0.00	0.04	-0.03
COUNT20	0.05	0.00	0.05	0.00	0.10	-0.07
COUNT10*	0.03	0.00	0.03	0.00	0.12	-0.05
NUMBER9	0.01	0.00	0.01	0.00	0.06	0.05
NUMBER23	0.02	0.00	0.02	0.00	0.01	-0.01
3RD LINE	0.03	0.00	0.03	0.01	0.17	-0.11
STICKBAT	0.02	0.00	0.02	0.01	0.17	-0.16
_789 10	0.01	0.00	0.02	0.00	0.09	0.07
51015_25	0.01	0.00	0.01	0.00	0.04	-0.02
3+2 CARS	0.02	0.00	0.02	0.00	0.05	-0.04
5-1ORANG	0.01	0.00	0.01	0.00	0.01	0.00
2+5MARBL	0.01	0.00	0.01	0.00	0.07	0.05
3+7PENNY	0.02	0.00	0.02	0.00	0.03	0.01
13___79	0.01	0.00	0.01	0.00	0.05	-0.04
COST$10	0.03	0.00	0.03	0.00	0.05	-0.04
8/2CANDY	0.03	0.00	0.03	0.00	0.04	0.01
15/5CARS	0.02	0.00	0.02	0.00	0.04	-0.01
2CRAYONS	0.01	0.00	0.01	0.00	0.03	0.01
3BANANAS	0.03	0.00	0.03	0.00	0.04	-0.02
6BANANAS	0.01	0.01	0.01	0.01	0.04	0.04
NUMBER 4	0.01	0.01	0.01	0.00	0.02	0.02
NUMBER 7	0.01	0.01	0.01	0.01	0.03	0.02
NUMBER17	0.02	0.00	0.02	0.01	0.03	-0.01
SQUARE	0.02	0.00	0.02	0.00	0.02	-0.01
LG-SM-SM	0.01	0.00	0.01	0.00	0.03	0.01
000X	0.01	0.00	0.01	0.00	0.05	-0.01
HALFOVAL	0.01	-0.01	0.01	0.00	0.05	-0.02
2+3STICK	0.01	0.00	0.01	0.00	0.05	0.04
3-1PENCL	0.01	0.00	0.01	0.01	0.05	-0.04
2+5CIRCL	0.01	0.00	0.01	0.00	0.03	-0.01
8-6CRAYN	0.01	0.00	0.01	0.00	0.06	0.05
PNTBRUSH	0.01	0.00	0.00	0.00	0.03	0.02
#CHOC	0.01	0.00	0.01	0.00	0.06	0.05
#VANILLA	0.01	0.00	0.01	0.00	0.07	0.05
#BUGS	0.01	0.00	0.01	0.00	0.04	0.00
4 LINES	0.01	-0.01	0.01	0.00	0.10	-0.08
SHAPES	0.01	0.00	0.01	0.00	0.08	0.01
PATTERN	0.01	0.00	0.00	0.00	0.09	-0.08
12 BY 2S	0.01	0.00	0.01	0.00	0.06	0.04
2+2	0.03	-0.01	0.03	-0.01	0.06	0.04
3+4	0.05	0.00	0.05	-0.01	0.16	0.10
1+7	0.01	0.01	0.01	0.01	0.14	0.12
3+3	0.01	-0.01	0.02	-0.01	0.09	0.05
11+3	0.03	0.00	0.03	0.00	0.16	0.14
12+6	0.04	0.00	0.04	0.00	0.13	0.12

Table D-5.—Root mean square deviation and mean deviation for English versus Spanish mathematics test items (continued)

Item	Total		English		Spanish	
	RMSD	MD	RMSD	MD	RMSD	MD
17-4	0.03	0.00	0.03	0.00	0.08	0.06
#STRAW	0.00	0.00	0.00	0.00	0.04	-0.01
#MORE	0.01	0.00	0.01	0.00	0.27	-0.26
HEADSUP	0.02	0.00	0.02	0.00	0.10	-0.09
HOWMANY$	0.01	0.00	0.01	0.00	0.05	0.02
25-14BKS	0.01	0.00	0.01	0.00	0.07	-0.06
12-? PEN	0.01	0.00	0.01	0.00	0.11	-0.09
GOALS	0.01	0.00	0.01	0.00	0.08	0.03
CHANGE	0.01	0.00	0.01	0.00	0.02	-0.02
17CENTS	0.01	0.00	0.01	0.00	0.07	0.01
BD CAKE	0.00	0.00	0.00	0.00	0.13	0.11
24/4 TAB	0.01	0.00	0.01	0.00	0.06	-0.01
2-1+2	0.02	0.00	0.02	0.00	0.05	0.01
9-2	0.03	-0.01	0.03	-0.01	0.10	0.07
7-3	0.03	-0.01	0.03	-0.01	0.13	0.12
4+4-2	0.01	0.00	0.01	0.00	0.03	-0.02
6+7	0.02	-0.01	0.02	-0.01	0.13	0.11
12-9	0.03	0.00	0.03	0.00	0.21	0.19
26+20	0.02	0.00	0.02	0.00	0.13	0.12
Average	0.02	0.00	0.02	0.00	0.07	0.01

Appendix E

Score Statistics for Indirect and Psychomotor Measures for Selected Subgroups

Table E-1.—Academic rating scale: language and literacy (range of possible values: 1-5)

Characteristic	Round 1			Round 2			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,688	2.57	0.79	18,908	3.33	0.81	14,530	3.40	0.93
Male	9,020	2.50	0.78	9,654	3.24	0.80	7,339	3.29	0.93
Female	8,667	2.64	0.78	9,253	3.43	0.80	7,191	3.51	0.92
White, non-Hispanic	9,637	2.73	0.75	10,638	3.46	0.78	8,389	3.52	0.89
Black, non-Hispanic	2,693	2.41	0.77	2,819	3.19	0.79	2,014	3.19	0.97
Hispanic, race specified	1,492	2.31	0.76	1,508	3.15	0.79	1,139	3.29	0.94
Hispanic, race not specified	1,678	2.22	0.75	1,711	3.06	0.83	1,273	3.13	0.95
Asian	1,170	2.50	0.86	1,193	3.35	0.80	915	3.54	0.90
Hawaiian, other Pacific Islander	200	2.38	0.89	199	3.20	0.74	173	2.97	0.84
American Indian, Alaska Native	317	2.31	0.75	335	2.99	0.78	257	2.98	0.88
More than one race, non-Hispanic	448	2.55	0.78	461	3.32	0.78	339	3.38	0.94
SES: first quintile	3,140	2.13	0.71	3,245	2.92	0.77	2,112	2.96	0.91
SES: second quintile	3,262	2.44	0.74	3,490	3.22	0.78	2,485	3.29	0.91
SES: third quintile	3,367	2.58	0.71	3,619	3.34	0.74	2,669	3.44	0.89
SES: fourth quintile	3,435	2.75	0.75	3,774	3.51	0.77	2,843	3.56	0.87
SES: fifth quintile	3,664	3.00	0.74	4,019	3.70	0.76	3,271	3.79	0.82
Public school	13,973	2.51	0.77	14,797	3.29	0.81	11,507	3.36	0.94
Private school	3,715	2.93	0.75	4,111	3.54	0.74	3,023	3.65	0.82

Table E-2.—Academic rating scale: mathematical thinking (range of possible values: 1-5)

	Round 1			Round 2			Round 4		
Characteristic	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	14,462	2.54	0.82	18,744	3.50	0.86	14,378	3.43	0.90
Male	7,369	2.50	0.82	9,563	3.45	0.87	7,264	3.42	0.92
Female	7,092	2.59	0.81	9,180	3.55	0.84	7,114	3.44	0.88
White, non-Hispanic	7,932	2.71	0.80	10,530	3.64	0.83	8,296	3.57	0.87
Black, non-Hispanic	2,154	2.33	0.77	2,790	3.29	0.85	1,992	3.16	0.93
Hispanic, race specified	1,244	2.30	0.76	1,502	3.29	0.84	1,126	3.29	0.91
Hispanic, race not specified	1,418	2.25	0.75	1,708	3.23	0.89	1,268	3.18	0.89
Asian	943	2.56	0.91	1,187	3.55	0.84	907	3.58	0.87
Hawaiian, other Pacific Islander	144	2.37	0.91	198	3.29	0.82	174	3.00	0.81
American Indian, Alaska Native	252	2.28	0.75	331	3.09	0.88	252	3.06	0.90
More than one race, non-Hispanic	339	2.53	0.82	454	3.51	0.81	332	3.45	0.86
SES: first quintile	2,577	2.15	0.70	3,234	3.07	0.85	2,101	3.01	0.88
SES: second quintile	2,692	2.40	0.76	3,477	3.39	0.85	2,463	3.30	0.89
SES: third quintile	2,748	2.55	0.75	3,577	3.51	0.81	2,640	3.46	0.87
SES: fourth quintile	2,789	2.71	0.81	3,719	3.68	0.80	2,810	3.59	0.84
SES: fifth quintile	2,986	2.97	0.82	3,979	3.88	0.77	3,223	3.82	0.79
Public school	11,567	2.49	0.79	14,733	3.46	0.86	11,393	3.39	0.91
Private school	2,895	2.89	0.85	4,011	3.72	0.80	2,985	3.68	0.78

Table E-3.—Academic rating scale: general knowledge (range of possible values: 1-5)

	Round 1			Round 2			Round 4		
Characteristic	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	15,263	2.64	1.00	18,828	3.55	0.99	13,863	3.26	0.99
Male	7,775	2.59	1.01	9,617	3.50	1.00	6,994	3.23	0.99
Female	7,487	2.69	1.00	9,210	3.61	0.98	6,869	3.30	0.98
White, non-Hispanic	8,298	2.85	1.01	10,588	3.72	0.96	7,970	3.40	0.96
Black, non-Hispanic	2,329	2.39	0.91	2,808	3.35	0.97	1,941	3.01	0.99
Hispanic, race specified	1,280	2.33	0.92	1,498	3.31	0.96	1,084	3.12	0.98
Hispanic, race not specified	1,480	2.28	0.90	1,707	3.26	1.02	1,223	2.99	0.99
Asian	994	2.52	1.06	1,190	3.44	1.00	870	3.33	0.99
Hawaiian, other Pacific Islander	178	2.40	0.94	198	3.23	1.02	172	2.82	0.88
American Indian, Alaska Native	274	2.34	0.92	339	3.00	0.99	247	2.87	0.94
More than one race, non-Hispanic	391	2.61	0.96	456	3.55	0.94	325	3.32	0.93
SES: first quintile	2,701	2.17	0.84	3,227	3.06	0.99	2,025	2.83	0.93
SES: second quintile	2,796	2.46	0.95	3,485	3.43	0.98	2,383	3.11	0.97
SES: third quintile	2,868	2.63	0.94	3,606	3.57	0.94	2,530	3.29	0.97
SES: fourth quintile	2,965	2.83	0.99	3,740	3.75	0.93	2,714	3.45	0.95
SES: fifth quintile	3,233	3.16	1.01	4,006	3.99	0.88	3,102	3.65	0.91
Public school	11,989	2.56	0.98	14,756	3.49	0.99	11,008	3.22	0.99
Private school	3,274	3.11	1.03	4,072	3.90	0.91	2,855	3.52	0.92

Table E-4.—Teacher rating: approaches to learning (range of possible values: 1-4)

		Round 1			Round 2			Round 4	
Characteristic	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,839	2.96	0.68	18,979	3.08	0.70	14,536	3.01	0.71
Male	9,610	2.82	0.68	9,686	2.94	0.70	7,343	2.87	0.71
Female	9,228	3.10	0.65	9,292	3.22	0.66	7,193	3.16	0.68
White, non-Hispanic	10,394	3.03	0.67	10,679	3.15	0.68	8,391	3.08	0.69
Black, non-Hispanic	2,817	2.78	0.71	2,826	2.88	0.74	2,013	2.77	0.75
Hispanic, race specified	1,558	2.91	0.66	1,513	3.03	0.68	1,140	3.00	0.70
Hispanic, race not specified	1,762	2.86	0.67	1,720	2.99	0.69	1,274	2.95	0.69
Asian	1,217	3.11	0.65	1,197	3.29	0.61	916	3.27	0.65
Hawaiian, other Pacific Islander	207	2.86	0.66	198	2.96	0.67	175	2.88	0.68
American Indian, Alaska Native	353	2.78	0.69	341	2.91	0.73	256	2.82	0.73
More than one race, non-Hispanic	478	2.92	0.66	461	3.07	0.68	340	3.02	0.69
SES: first quintile	3,309	2.73	0.70	3,253	2.84	0.72	2,116	2.77	0.74
SES: second quintile	3,481	2.89	0.68	3,499	3.01	0.70	2,491	2.93	0.73
SES: third quintile	3,586	2.98	0.65	3,628	3.09	0.69	2,676	3.02	0.70
SES: fourth quintile	3,699	3.05	0.67	3,789	3.18	0.68	2,839	3.10	0.67
SES: fifth quintile	3,901	3.16	0.62	4,040	3.28	0.61	3,268	3.24	0.62
Public school	14,898	2.94	0.69	14,841	3.06	0.71	11,521	2.99	0.72
Private school	3,941	3.04	0.63	4,138	3.17	0.64	3,015	3.13	0.64

Table E-5.—Teacher rating: self-control (range of possible values: 1-4)

	Round 1			Round 2			Round 4		
Characteristic	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,135	3.07	0.62	18,847	3.15	0.63	14,425	3.16	0.62
Male	9,254	2.97	0.63	9,621	3.05	0.65	7,289	3.06	0.63
Female	8,880	3.18	0.59	9,225	3.26	0.60	7,136	3.26	0.59
White, non-Hispanic	10,080	3.13	0.62	10,624	3.22	0.61	8,354	3.21	0.60
Black, non-Hispanic	2,723	2.89	0.64	2,808	2.94	0.68	1,990	2.94	0.67
Hispanic, race specified	1,462	3.06	0.60	1,501	3.12	0.61	1,122	3.17	0.61
Hispanic, race not specified	1,657	3.04	0.59	1,696	3.13	0.62	1,266	3.15	0.59
Asian	1,153	3.16	0.58	1,180	3.30	0.57	898	3.33	0.53
Hawaiian, other Pacific Islander	192	2.97	0.58	198	3.01	0.61	175	3.03	0.63
American Indian, Alaska Native	351	2.92	0.58	338	2.99	0.62	254	2.99	0.61
More than one race, non-Hispanic	465	3.03	0.63	458	3.14	0.62	336	3.19	0.63
SES: first quintile	3,116	2.94	0.64	3,219	3.02	0.64	2,090	3.01	0.65
SES: second quintile	3,364	3.03	0.62	3,475	3.11	0.63	2,478	3.12	0.64
SES: third quintile	3,481	3.08	0.62	3,612	3.15	0.64	2,656	3.16	0.61
SES: fourth quintile	3,576	3.13	0.61	3,767	3.21	0.63	2,819	3.20	0.61
SES: fifth quintile	3,779	3.19	0.59	4,012	3.28	0.59	3,250	3.32	0.56
Public school	14,337	3.07	0.62	14,723	3.15	0.63	11,433	3.16	0.62
Private school	3,798	3.07	0.60	4,124	3.17	0.63	2,992	3.18	0.59

Table E-6.—Teacher rating: interpersonal (range of possible values: 1-4)

	Round 1			Round 2			Round 4		
Characteristic	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,923	2.96	0.63	18,767	3.09	0.65	14,387	3.09	0.64
Male	9,082	2.85	0.63	9,554	2.98	0.65	7,252	2.97	0.64
Female	8,840	3.08	0.62	9,212	3.20	0.62	7,135	3.22	0.62
White, non-Hispanic	9,996	3.03	0.63	10,586	3.16	0.63	8,333	3.14	0.63
Black, non-Hispanic	2,685	2.81	0.64	2,789	2.91	0.68	1,986	2.90	0.68
Hispanic, race specified	1,441	2.94	0.62	1,483	3.05	0.64	1,118	3.10	0.62
Hispanic, race not specified	1,644	2.88	0.61	1,694	3.04	0.63	1,258	3.07	0.61
Asian	1,110	2.97	0.63	1,180	3.19	0.61	898	3.18	0.60
Hawaiian, other Pacific Islander	195	2.79	0.61	197	2.90	0.63	175	2.95	0.62
American Indian, Alaska Native	341	2.84	0.58	337	2.94	0.64	252	2.89	0.60
More than one race, non-Hispanic	460	2.94	0.62	457	3.07	0.63	337	3.10	0.64
SES: first quintile	3,056	2.80	0.64	3,200	2.93	0.65	2,082	2.91	0.66
SES: second quintile	3,307	2.91	0.64	3,451	3.04	0.64	2,459	3.04	0.65
SES: third quintile	3,445	2.99	0.62	3,598	3.10	0.65	2,648	3.11	0.64
SES: fourth quintile	3,544	3.03	0.62	3,759	3.16	0.65	2,817	3.15	0.63
SES: fifth quintile	3,769	3.11	0.61	4,003	3.24	0.61	3,247	3.25	0.60
Public school	14,101	2.95	0.64	14,678	3.08	0.65	11,394	3.08	0.65
Private school	3,822	3.01	0.61	4,089	3.14	0.63	2,993	3.15	0.61

Table E-7.—Teacher rating: externalizing problem behaviors (range of possible values: 1-4)

Characteristic	Round 1 N	Round 1 Mean	Round 1 S.D.	Round 2 N	Round 2 Mean	Round 2 S.D.	Round 4 N	Round 4 Mean	Round 4 S.D.
Total sample	18,609	1.64	0.65	18,907	1.69	0.66	14,448	1.67	0.65
Male	9,492	1.77	0.69	9,647	1.82	0.70	7,305	1.80	0.69
Female	9,116	1.51	0.57	9,259	1.55	0.58	7,143	1.54	0.57
White, non-Hispanic	10,293	1.61	0.63	10,649	1.64	0.63	8,349	1.63	0.63
Black, non-Hispanic	2,785	1.81	0.71	2,816	1.90	0.75	2,005	1.89	0.74
Hispanic, race specified	1,530	1.60	0.62	1,505	1.67	0.62	1,133	1.61	0.59
Hispanic, race not specified	1,720	1.61	0.63	1,704	1.65	0.62	1,263	1.62	0.62
Asian	1,199	1.47	0.55	1,191	1.48	0.54	899	1.49	0.54
Hawaiian, other Pacific Islander	204	1.75	0.69	197	1.84	0.69	174	1.75	0.60
American Indian, Alaska Native	349	1.80	0.61	341	1.83	0.62	254	1.77	0.63
More than one race, non-Hispanic	477	1.65	0.67	460	1.71	0.65	340	1.68	0.65
SES: first quintile	3,244	1.73	0.70	3,240	1.78	0.71	2,098	1.80	0.71
SES: second quintile	3,437	1.68	0.66	3,476	1.72	0.67	2,482	1.70	0.67
SES: third quintile	3,549	1.63	0.64	3,619	1.69	0.66	2,664	1.67	0.65
SES: fourth quintile	3,661	1.61	0.63	3,777	1.65	0.64	2,823	1.63	0.62
SES: fifth quintile	3,868	1.56	0.58	4,027	1.58	0.58	3,245	1.53	0.56
Public school	14,685	1.64	0.66	14,771	1.69	0.66	11,442	1.68	0.66
Private school	3,924	1.65	0.61	4,136	1.68	0.62	3,006	1.62	0.58

Table E-8.—Teacher rating: internalizing problem behaviors (range of possible values: 1-4)

Characteristic	Round 1 N	Round 1 Mean	Round 1 S.D.	Round 2 N	Round 2 Mean	Round 2 S.D.	Round 4 N	Round 4 Mean	Round 4 S.D.
Total sample	18,356	1.55	0.53	18,806	1.59	0.53	14,362	1.61	0.53
Male	9,344	1.57	0.54	9,594	1.60	0.54	7,254	1.62	0.54
Female	9,011	1.54	0.53	9,211	1.58	0.52	7,108	1.59	0.53
White, non-Hispanic	10,208	1.54	0.53	10,642	1.58	0.52	8,314	1.60	0.52
Black, non-Hispanic	2,708	1.58	0.56	2,782	1.63	0.58	1,980	1.65	0.57
Hispanic, race specified	1,500	1.55	0.52	1,488	1.63	0.52	1,127	1.58	0.54
Hispanic, race not specified	1,703	1.58	0.55	1,685	1.59	0.50	1,256	1.61	0.53
Asian	1,161	1.49	0.48	1,180	1.49	0.47	895	1.49	0.46
Hawaiian, other Pacific Islander	199	1.60	0.56	193	1.69	0.51	173	1.54	0.50
American Indian, Alaska Native	348	1.66	0.57	337	1.66	0.48	248	1.74	0.61
More than one race, non-Hispanic	476	1.57	0.54	456	1.64	0.55	338	1.69	0.55
SES: first quintile	3,185	1.65	0.59	3,200	1.67	0.56	2,080	1.73	0.59
SES: second quintile	3,386	1.59	0.57	3,467	1.63	0.55	2,451	1.66	0.56
SES: third quintile	3,519	1.53	0.51	3,595	1.59	0.53	2,656	1.58	0.51
SES: fourth quintile	3,619	1.50	0.50	3,761	1.54	0.51	2,808	1.57	0.49
SES: fifth quintile	3,820	1.49	0.47	4,028	1.51	0.47	3,238	1.51	0.47
Public school	14,457	1.56	0.54	14,684	1.60	0.54	11,378	1.61	0.54
Private school	3,899	1.52	0.49	4,122	1.55	0.48	2,984	1.56	0.48

Table E-9.—Parent rating: approaches to learning (range of possible values: 1-4)

Characteristic	Round 1			Round 2			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,521	3.11	0.48	18,253	3.12	0.48	14,990	3.09	0.50
Male	8,896	3.06	0.49	9,336	3.06	0.48	7,659	3.03	0.49
Female	8,625	3.17	0.47	8,917	3.19	0.47	7,331	3.15	0.49
White, non-Hispanic	10,016	3.16	0.46	10,538	3.15	0.46	8,923	3.12	0.48
Black, non-Hispanic	2,532	3.07	0.51	2,591	3.08	0.51	1,969	3.07	0.53
Hispanic, race specified	1,508	3.05	0.49	1,559	3.08	0.50	1,208	3.04	0.51
Hispanic, race not specified	1,557	3.01	0.49	1,594	3.07	0.51	1,270	2.99	0.51
Asian	925	3.04	0.50	1,003	3.06	0.52	797	3.04	0.54
Hawaiian, other Pacific Islander	195	2.91	0.44	194	2.97	0.52	161	3.01	0.47
American Indian, Alaska Native	298	3.08	0.52	298	3.14	0.49	276	3.14	0.49
More than one race, non-Hispanic	466	3.17	0.48	452	3.16	0.49	368	3.08	0.50
SES: first quintile	3,127	2.96	0.52	3,237	3.00	0.52	2,472	2.95	0.55
SES: second quintile	3,345	3.09	0.49	3,446	3.09	0.49	2,784	3.05	0.49
SES: third quintile	3,518	3.13	0.47	3,611	3.14	0.48	2,967	3.11	0.50
SES: fourth quintile	3,655	3.17	0.44	3,841	3.17	0.45	3,139	3.12	0.47
SES: fifth quintile	3,876	3.22	0.44	4,118	3.22	0.44	3,618	3.19	0.44
Public school	13,770	3.10	0.49	14,075	3.11	0.49	11,845	3.08	0.50
Private school	3,751	3.18	0.44	4,178	3.18	0.45	3,084	3.17	0.48

Table E-10.—Parent rating: self-control (range of possible values: 1-4)

Characteristic	Round 1 N	Round 1 Mean	Round 1 S.D.	Round 2 N	Round 2 Mean	Round 2 S.D.	Round 4 N	Round 4 Mean	Round 4 S.D.
Total sample	17,515	2.83	0.52	18,252	2.87	0.52	14,989	2.94	0.51
Male	8,895	2.80	0.52	9,336	2.84	0.52	7,659	2.91	0.52
Female	8,620	2.86	0.52	8,916	2.90	0.52	7,330	2.98	0.50
White, non-Hispanic	10,016	2.83	0.49	10,538	2.88	0.48	8,922	2.95	0.49
Black, non-Hispanic	2,530	2.82	0.59	2,592	2.85	0.59	1,969	2.94	0.57
Hispanic, race specified	1,508	2.82	0.54	1,559	2.87	0.57	1,209	2.92	0.54
Hispanic, race not specified	1,555	2.83	0.54	1,592	2.82	0.57	1,270	2.92	0.54
Asian	923	2.94	0.49	1,004	2.99	0.46	796	3.06	0.50
Hawaiian, other Pacific Islander	195	2.75	0.44	193	2.80	0.44	161	2.80	0.44
American Indian, Alaska Native	298	2.81	0.55	298	2.91	0.53	276	2.90	0.62
More than one race, non-Hispanic	466	2.78	0.53	452	2.86	0.51	368	2.89	0.48
SES: first quintile	3,125	2.68	0.60	3,240	2.71	0.63	2,471	2.79	0.59
SES: second quintile	3,345	2.79	0.54	3,443	2.83	0.52	2,786	2.87	0.55
SES: third quintile	3,515	2.85	0.50	3,611	2.90	0.48	2,967	2.98	0.51
SES: fourth quintile	3,654	2.89	0.48	3,841	2.94	0.47	3,138	3.01	0.45
SES: fifth quintile	3,876	2.93	0.45	4,117	2.97	0.43	3,617	3.04	0.43
Public school	13,765	2.81	0.53	14,076	2.85	0.53	11,844	2.93	0.52
Private school	3,750	2.91	0.46	4,176	2.96	0.44	3,084	3.06	0.45

Table E-11.—Parent rating: social interaction (range of possible values: 1-4)

	Round 1			Round 2			Round 4		
Characteristic	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,517	3.32	0.56	18,271	3.42	0.53	14,997	3.39	0.55
Male	8,895	3.29	0.57	9,347	3.38	0.54	7,659	3.34	0.56
Female	8,622	3.36	0.55	8,924	3.47	0.51	7,338	3.45	0.53
White, non-Hispanic	10,016	3.38	0.53	10,546	3.47	0.50	8,923	3.45	0.51
Black, non-Hispanic	2,531	3.32	0.57	2,593	3.39	0.54	1,971	3.37	0.56
Hispanic, race specified	1,508	3.20	0.59	1,560	3.32	0.57	1,208	3.27	0.60
Hispanic, race not specified	1,556	3.14	0.62	1,595	3.30	0.61	1,269	3.21	0.61
Asian	924	3.10	0.59	1,010	3.25	0.58	803	3.20	0.61
Hawaiian, other Pacific Islander	195	3.00	0.63	193	3.18	0.67	161	3.15	0.57
American Indian, Alaska Native	298	3.30	0.58	298	3.48	0.50	276	3.40	0.56
More than one race, non-Hispanic	465	3.39	0.51	452	3.44	0.52	368	3.43	0.54
SES: first quintile	3,127	3.15	0.62	3,243	3.26	0.60	2,473	3.20	0.62
SES: second quintile	3,343	3.31	0.56	3,447	3.42	0.53	2,786	3.36	0.56
SES: third quintile	3,516	3.36	0.54	3,611	3.47	0.50	2,969	3.45	0.51
SES: fourth quintile	3,655	3.40	0.52	3,849	3.49	0.49	3,140	3.47	0.51
SES: fifth quintile	3,876	3.38	0.53	4,121	3.46	0.50	3,619	3.47	0.50
Public school	13,767	3.31	0.57	14,091	3.41	0.54	11,850	3.38	0.55
Private school	3,750	3.39	0.53	4,180	3.48	0.50	3,086	3.46	0.50

Table E-12.—Parent rating: sad/lonely (range of possible values: 1-4)

Characteristic	Round 1			Round 2			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,503	1.54	0.41	18,233	1.55	0.40	14,985	1.54	0.41
Male	8,887	1.54	0.41	9,325	1.55	0.41	7,654	1.54	0.42
Female	8,616	1.54	0.40	8,908	1.55	0.40	7,331	1.54	0.40
White, non-Hispanic	10,009	1.54	0.38	10,533	1.55	0.38	8,919	1.54	0.39
Black, non-Hispanic	2,531	1.56	0.47	2,588	1.56	0.46	1,970	1.56	0.45
Hispanic, race specified	1,507	1.52	0.40	1,556	1.53	0.43	1,207	1.53	0.42
Hispanic, race not specified	1,557	1.51	0.43	1,591	1.51	0.41	1,271	1.53	0.44
Asian	916	1.61	0.42	997	1.61	0.40	796	1.58	0.38
Hawaiian, other Pacific Islander	195	1.82	0.47	194	1.74	0.47	161	1.71	0.38
American Indian, Alaska Native	298	1.60	0.42	298	1.59	0.43	276	1.54	0.46
More than one race, non-Hispanic	466	1.59	0.44	452	1.57	0.38	367	1.55	0.43
SES: first quintile	3,123	1.59	0.47	3,226	1.59	0.48	2,471	1.60	0.47
SES: second quintile	3,340	1.55	0.44	3,442	1.55	0.41	2,782	1.55	0.43
SES: third quintile	3,515	1.53	0.38	3,609	1.54	0.39	2,967	1.54	0.40
SES: fourth quintile	3,652	1.52	0.37	3,839	1.53	0.37	3,138	1.52	0.38
SES: fifth quintile	3,873	1.53	0.36	4,117	1.55	0.35	3,617	1.51	0.35
Public school	13,754	1.54	0.41	14,058	1.55	0.41	11,841	1.55	0.41
Private school	3,749	1.53	0.36	4,175	1.54	0.36	3,083	1.50	0.35

Table E-13.—Parent rating: impulsive/overactive (range of possible values: 1-4)

Characteristic	Round 1			Round 2			Round 4		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	17,406	1.98	0.69	18,092	1.96	0.70	14,909	1.89	0.69
Male	8,843	2.05	0.72	9,257	2.04	0.72	7,621	1.97	0.72
Female	8,563	1.89	0.65	8,835	1.88	0.67	7,288	1.80	0.64
White, non-Hispanic	9,973	1.92	0.66	10,481	1.91	0.66	8,902	1.85	0.66
Black, non-Hispanic	2,518	2.18	0.76	2,575	2.13	0.77	1,962	2.07	0.77
Hispanic, race specified	1,503	1.98	0.70	1,550	1.97	0.73	1,202	1.88	0.70
Hispanic, race not specified	1,538	1.95	0.71	1,571	2.02	0.74	1,258	1.87	0.67
Asian	896	1.89	0.63	955	1.87	0.64	767	1.78	0.66
Hawaiian, other Pacific Islander	195	2.07	0.69	188	2.07	0.69	159	2.07	0.72
American Indian, Alaska Native	297	2.04	0.73	297	1.97	0.72	276	1.98	0.71
More than one race, non-Hispanic	462	2.08	0.73	451	2.02	0.74	366	1.96	0.74
SES: first quintile	3,094	2.17	0.78	3,186	2.17	0.78	2,442	2.07	0.76
SES: second quintile	3,320	2.06	0.74	3,405	2.04	0.73	2,762	1.98	0.72
SES: third quintile	3,498	1.98	0.67	3,589	1.95	0.68	2,961	1.87	0.69
SES: fourth quintile	3,638	1.89	0.63	3,820	1.89	0.63	3,129	1.84	0.65
SES: fifth quintile	3,856	1.79	0.58	4,092	1.78	0.59	3,605	1.72	0.57
Public school	13,671	2.00	0.70	13,926	1.99	0.71	11,772	1.91	0.70
Private school	3,735	1.85	0.62	4,166	1.84	0.63	3,076	1.74	0.61

Table E-14.—Psychomotor: fine motor skills, gross motor skills, composite motor skills

Characteristic	Fine motor skills (range of possible values 0-9)			Gross motor skills (range of possible values 0-8)			Composite motor skills (range of possible values 0-17)		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
Total sample	18,559	5.75	2.06	18,493	6.32	1.86	18,422	12.08	3.10
Male	9,440	5.60	2.08	9,402	6.10	1.96	9,364	11.71	3.19
Female	9,118	5.90	2.03	9,090	6.57	1.73	9,057	12.47	2.96
White, non-Hispanic	10,378	5.97	1.97	10,372	6.27	1.89	10,322	12.25	3.07
Black, non-Hispanic	2,845	5.04	2.19	2,832	6.59	1.73	2,826	11.64	3.13
Hispanic, race specified	1,589	5.71	2.07	1,575	6.39	1.90	1,572	12.11	3.19
Hispanic, race not specified	1,792	5.48	2.09	1,768	6.12	1.92	1,763	11.62	3.16
Asian	895	6.59	1.88	892	6.35	1.78	889	12.94	2.85
Hawaiian, other Pacific Islander	187	5.97	1.95	185	6.25	1.83	185	12.20	3.13
American Indian, Alaska Native	352	5.67	1.91	348	6.39	1.74	346	12.04	2.91
More than one race, non-Hispanic	473	5.65	1.96	473	6.42	1.78	471	12.07	2.96
SES: first quintile	3,252	4.98	2.16	3,220	6.17	1.97	3,206	11.17	3.29
SES: second quintile	3,419	5.48	2.10	3,403	6.20	1.93	3,393	11.69	3.19
SES: third quintile	3,541	5.86	1.98	3,522	6.39	1.80	3,517	12.25	2.97
SES: fourth quintile	3,659	6.10	1.90	3,661	6.42	1.79	3,641	12.53	2.90
SES: fifth quintile	3,870	6.37	1.84	3,873	6.47	1.78	3,854	12.84	2.82
Public school	14,637	5.66	2.07	14,580	6.31	1.87	14,518	11.97	3.12
Private school	3,922	6.26	1.91	3,913	6.44	1.81	3,904	12.70	2.92

Listing of NCES Working Papers to Date

Working papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/). You can also contact Sheilah Jupiter at (202) 502–7444 (sheilah_jupiter@ed.gov) if you are interested in any of the following papers.

Listing of NCES Working Papers by Program Area

No.	Title	NCES contact
Baccalaureate and Beyond (B&B)		
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
Beginning Postsecondary Students (BPS) Longitudinal Study		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
2001–04	Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
Common Core of Data (CCD)		
95–12	Rural Education Data User's Guide	Samuel Peng
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97–15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
2000–12	Coverage Evaluation of the 1994–95 Common Core of Data: Public Elementary/Secondary School Universe Survey	Beth Young
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2002–02	School Locale Codes 1987 - 2000	Frank Johnson
Data Development		
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Decennial Census School District Project		
95–12	Rural Education Data User's Guide	Samuel Peng
96–04	Census Mapping Project/School District Data Book	Tai Phan
98–07	Decennial Census School District Project Planning Report	Tai Phan
Early Childhood Longitudinal Study (ECLS)		
96–08	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96–18	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
97–24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97–36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
2001–03	Measures of Socio-Emotional Development in Middle Childhood	Elvira Hausken

No.	Title	NCES contact
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
2002-05	Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade	Elvira Hausken
Education Finance Statistics Center (EDFIN)		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
High School and Beyond (HS&B)		
95–12	Rural Education Data User's Guide	Samuel Peng
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
HS Transcript Studies		
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
International Adult Literacy Survey (IALS)		
97–33	Adult Literacy: An International Perspective	Marilyn Binkley
Integrated Postsecondary Education Data System (IPEDS)		
97–27	Pilot Test of IPEDS Finance Survey	Peter Stowe
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–14	IPEDS Finance Data Comparisons Under the 1997 Financial Accounting Standards for Private, Not-for-Profit Institutes: A Concept Paper	Peter Stowe
National Assessment of Adult Literacy (NAAL)		
98–17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999–09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999–09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999–09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999–09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999–09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999–09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999–09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
2000–05	Secondary Statistical Modeling With the National Assessment of Adult Literacy: Implications for the Design of the Background Questionnaire	Sheida White
2000–06	Using Telephone and Mail Surveys as a Supplement or Alternative to Door-to-Door Surveys in the Assessment of Adult Literacy	Sheida White
2000–07	"How Much Literacy is Enough?" Issues in Defining and Reporting Performance Standards for the National Assessment of Adult Literacy	Sheida White
2000–08	Evaluation of the 1992 NALS Background Survey Questionnaire: An Analysis of Uses with Recommendations for Revisions	Sheida White
2000–09	Demographic Changes and Literacy Development in a Decade	Sheida White
2001–08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
National Assessment of Educational Progress (NAEP)		
95–12	Rural Education Data User's Guide	Samuel Peng
97–29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97–30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman

BEST COPY AVAILABLE

No.	Title	NCES contact
97–31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman
97–32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires)	Steven Gorman
97–37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Steven Gorman
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
2001–19	The Measurement of Home Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Graders to Questionnaire Items and Parental Assessment of the Invasiveness of These Items	Arnold Goldstein
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
National Education Longitudinal Study of 1988 (NELS:88)		
95–04	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95–05	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings
95–06	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95–07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95–12	Rural Education Data User's Guide	Samuel Peng
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98–06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
2001–16	Imputation of Test Scores in the National Education Longitudinal Study of 1988	Ralph Lee
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
National Household Education Survey (NHES)		
95–12	Rural Education Data User's Guide	Samuel Peng
96–13	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96–14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–21	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96–29	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler

No.	Title	NCES contact
96–30	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97–02	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–03	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97–04	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–05	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–06	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97–08	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler
97–19	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97–20	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97–28	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97–34	Comparison of Estimates from the 1993 National Household Education Survey	Kathryn Chandler
97–35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
97–38	Reinterview Results for the Parent and Youth Components of the 1996 National Household Education Survey	Kathryn Chandler
97–39	Undercoverage Bias in Estimates of Characteristics of Households and Adults in the 1996 National Household Education Survey	Kathryn Chandler
97–40	Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey	Kathryn Chandler
98–03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
National Longitudinal Study of the High School Class of 1972 (NLS-72)		
95–12	Rural Education Data User's Guide	Samuel Peng
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
National Postsecondary Student Aid Study (NPSAS)		
96–17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
2000–17	National Postsecondary Student Aid Study:2000 Field Test Methodology Report	Andrew G. Malizio
2002–03	National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000), CATI Nonresponse Bias Analysis Report.	Andrew Malizio
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
National Study of Postsecondary Faculty (NSOPF)		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
Postsecondary Education Descriptive Analysis Reports (PEDAR)		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Private School Universe Survey (PSS)		
95–16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95–17	Estimates of Expenditures for Private K–12 Schools	Stephen Broughman
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96–26	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman

No.	Title	NCES contact
96–27	Intersurvey Consistency in NCES Private School Surveys for 1993–94	Steven Kaufman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Recent College Graduates (RCG)		
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
Schools and Staffing Survey (SASS)		
94–01	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94–02	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94–03	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94–04	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94–06	Six Papers on Teachers from the 1990–91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
95–01	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95–02	QED Estimates of the 1990–91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95–03	Schools and Staffing Survey: 1990–91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95–08	CCD Adjustment to the 1990–91 SASS: A Comparison of Estimates	Dan Kasprzyk
95–09	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95–10	The Results of the 1991–92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95–12	Rural Education Data User's Guide	Samuel Peng
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95–15	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95–16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95–18	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96–01	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk
96–02	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96–05	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96–06	The Schools and Staffing Survey (SASS) for 1998–99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96–07	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96–09	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998–99 SASS	Dan Kasprzyk
96–10	1998–99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96–11	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96–12	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96–15	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96–23	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96–24	National Assessments of Teacher Quality	Dan Kasprzyk
96–25	Measures of Inservice Professional Development: Suggested Items for the 1998–1999 Schools and Staffing Survey	Dan Kasprzyk

No.	Title	NCES contact
96–28	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
97–01	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97–10	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993–94 School Year	Dan Kasprzyk
97–11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97–12	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97–14	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97–18	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
97–23	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97–41	Selected Papers on the Schools and Staffing Survey: Papers Presented at the 1997 Meeting of the American Statistical Association	Steve Kaufman
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98–02	Response Variance in the 1993–94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
98–05	SASS Documentation: 1993–94 SASS Student Sampling Problems; Solutions for Determining the Numerators for the SASS Private School (3B) Second-Stage Factors	Steven Kaufman
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
98–12	A Bootstrap Variance Estimator for Systematic PPS Sampling	Steven Kaufman
98–13	Response Variance in the 1994–95 Teacher Follow-up Survey	Steven Kaufman
98–14	Variance Estimation of Imputed Survey Data	Steven Kaufman
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
98–16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999–02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Fieldtest Results to Improve Item Construction	Dan Kasprzyk
1999–10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
1999–12	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume III: Public-Use Codebook	Kerry Gruber
1999–13	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
1999–14	1994–95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
1999–17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
Third International Mathematics and Science Study (TIMSS)		
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein

No.	Title	NCES contact
2002–01	Legal and Ethical Issues in the Use of Video in Education Research	Patrick Gonzales

Listing of NCES Working Papers by Subject

No.	Title	NCES contact
Achievement (student) – mathematics		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
Adult education		
96–14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
98–03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Adult literacy—see Literacy of adults		
American Indian – education		
1999–13	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
Assessment/achievement		
95–12	Rural Education Data User's Guide	Samuel Peng
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
97–29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Larry Ogle
97–30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Larry Ogle
97–31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Larry Ogle
97–32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questions)	Larry Ogle
97–37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Larry Ogle
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
2001–19	The Measurement of Home Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Graders to Questionnaire Items and Parental Assessment of the Invasiveness of These Items	Arnold Goldstein
2002-05	Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade	Elvira Hausken
Beginning students in postsecondary education		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2001–04	Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper

No.	Title	NCES contact
Civic participation		
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Climate of schools		
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
Cost of education indices		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Course-taking		
95–12	Rural Education Data User's Guide	Samuel Peng
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
Crime		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Curriculum		
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Customer service		
1999–10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Data quality		
97–13	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
2001–19	The Measurement of Home Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Graders to Questionnaire Items and Parental Assessment of the Invasiveness of These Items	Arnold Goldstein
Data warehouse		
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Design effects		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
Dropout rates, high school		
95–07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
Early childhood education		
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler

No.	Title	NCES contact
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
97–24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97–36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
2001–03	Measures of Socio-Emotional Development in Middle School	Elvira Hausken
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
2002-05	Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade	Elvira Hausken
Educational attainment		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Educational research		
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2002–01	Legal and Ethical Issues in the Use of Video in Education Research	Patrick Gonzales
Eighth-graders		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
Employment		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
Employment – after college		
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Engineering		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Enrollment – after college		
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Faculty – higher education		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Fathers – role in education		
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
Finance – elementary and secondary schools		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman

No.	Title	NCES contact
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Finance – postsecondary		
97–27	Pilot Test of IPEDS Finance Survey	Peter Stowe
2000–14	IPEDS Finance Data Comparisons Under the 1997 Financial Accounting Standards for Private, Not-for-Profit Institutes: A Concept Paper	Peter Stowe
Finance – private schools		
95–17	Estimates of Expenditures for Private K–12 Schools	Stephen Broughman
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Geography		
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
Graduate students		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Graduates of postsecondary education		
2001–15	Baccalaureate and Beyond Longitudinal Study: 2000/01 Follow-Up Field Test Methodology Report	Andrew G. Malizio
Imputation		
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meeting	Dan Kasprzyk
2001–10	Comparison of Proc Impute and Schafer's Multiple Imputation Software	Sam Peng
2001–16	Imputation of Test Scores in the National Education Longitudinal Study of 1988	Ralph Lee
2001–17	A Study of Imputation Algorithms	Ralph Lee
2001–18	A Study of Variance Estimation Methods	Ralph Lee
Inflation		
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
Institution data		
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Instructional resources and practices		
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
International comparisons		
97–11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97–16	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97–17	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
International comparisons – math and science achievement		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales

No.	Title	NCES contact
Libraries		
94–07	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Limited English Proficiency		
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Literacy of adults		
98–17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999–09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999–09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999–09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999–09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999–09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999–09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999–09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–05	Secondary Statistical Modeling With the National Assessment of Adult Literacy: Implications for the Design of the Background Questionnaire	Sheida White
2000–06	Using Telephone and Mail Surveys as a Supplement or Alternative to Door-to-Door Surveys in the Assessment of Adult Literacy	Sheida White
2000–07	"How Much Literacy is Enough?" Issues in Defining and Reporting Performance Standards for the National Assessment of Adult Literacy	Sheida White
2000–08	Evaluation of the 1992 NALS Background Survey Questionnaire: An Analysis of Uses with Recommendations for Revisions	Sheida White
2000–09	Demographic Changes and Literacy Development in a Decade	Sheida White
2001–08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
Literacy of adults – international		
97–33	Adult Literacy: An International Perspective	Marilyn Binkley
Mathematics		
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
Parental involvement in education		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West

BEST COPY AVAILABLE

No.	Title	NCES contact
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
2001–19	The Measurement of Home Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Graders to Questionnaire Items and Parental Assessment of the Invasiveness of These Items	Arnold Goldstein
Participation rates		
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
Postsecondary education		
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Postsecondary education – persistence and attainment		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Postsecondary education – staff		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Principals		
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Private schools		
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Projections of education statistics		
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Public school finance		
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Public schools		
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999–02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
2000–12	Coverage Evaluation of the 1994–95 Public Elementary/Secondary School Universe Survey	Beth Young
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2002–02	Locale Codes 1987 - 2000	Frank Johnson

No.	Title	NCES contact
Public schools – secondary		
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Reform, educational		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
Response rates		
98–02	Response Variance in the 1993–94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
School districts		
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
School districts, public		
98–07	Decennial Census School District Project Planning Report	Tai Phan
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
School districts, public – demographics of		
96–04	Census Mapping Project/School District Data Book	Tai Phan
Schools		
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
2002–02	Locale Codes 1987 - 2000	Frank Johnson
Schools – safety and discipline		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Science		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
Software evaluation		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
Staff		
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Staff – higher education institutions		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
Staff – nonprofessional		
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
State		
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young

No.	Title	NCES contact
Statistical methodology		
97–21	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
Statistical standards and methodology		
2001–05	Using TIMSS to Analyze Correlates of Performance Variation in Mathematics	Patrick Gonzales
2002–04	Improving Consistency of Response Categories Across NCES Surveys	Marilyn Seastrom
Students with disabilities		
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
Survey methodology		
96–17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
97–15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97–35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
98–06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
98–16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–12	Coverage Evaluation of the 1994–95 Public Elementary/Secondary School Universe Survey	Beth Young
2000–17	National Postsecondary Student Aid Study:2000 Field Test Methodology Report	Andrew G. Malizio
2001–04	Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
2001–07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001–11	Impact of Selected Background Variables on Students' NAEP Math Performance	Arnold Goldstein
2001–13	The Effects of Accommodations on the Assessment of LEP Students in NAEP	Arnold Goldstein
2001–19	The Measurement of Home Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Graders to Questionnaire Items and Parental Assessment of the Invasiveness of These Items	Arnold Goldstein
2002–01	Legal and Ethical Issues in the Use of Video in Education Research	Patrick Gonzales
2002–02	Locale Codes 1987 - 2000	Frank Johnson
2002–03	National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000), CATI Nonresponse Bias Analysis Report.	Andrew Malizio
Teachers		
98–13	Response Variance in the 1994–95 Teacher Follow-up Survey	Steven Kaufman
1999–14	1994–95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Teachers – instructional practices of		
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Teachers – opinions regarding safety		
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Teachers – performance evaluations		
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk

BEST COPY AVAILABLE

No.	Title	NCES contact
Teachers – qualifications of		
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
Teachers – salaries of		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Training		
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Variance estimation		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2001–18	A Study of Variance Estimation Methods	Ralph Lee
Violence		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Vocational education		
95–12	Rural Education Data User's Guide	Samuel Peng
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (1/2003)